

EVOLUTIONARY GENETICS OF THE INNATE IMMUNE SYSTEM IN THE
GENUS *DROSOPHILA*

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Timothy Bradford Sackton

August 2008

© 2008 Timothy Bradford Sackton

EVOLUTIONARY GENETICS OF THE INNATE IMMUNE SYSTEM IN THE
GENUS *DROSOPHILA*

Timothy Bradford Sackton, Ph. D.

Cornell University 2008

Host-parasite interactions are ubiquitous in nature: all organisms, from bacteria to humans, must defend themselves against parasites and pathogens that seek to use them as a substrate for growth, reproduction, and survival. This coevolutionary dynamic has important implications for genetic architecture of host defense genes, as well as the molecular evolution of pathogenic defense systems. In the first part of my dissertation, I focus on characterizing the underlying genetic architecture of variation for *Drosophila* immune function. Using a candidate gene association testing approach, I identify genetic markers in several genes (principally encoding recognition proteins) that statistically associate with differences in resistance to bacterial infection. I also characterize the transcriptional response to infection at over 300 infection-regulated genes, and show that differences in the overall level of transcription of genes encoding effector proteins significantly correlates with differences in bacterial load. Understanding the genetic architecture of immune function provides insight into the puzzle of why extensive genetic variation for an important fitness trait persists in natural populations.

In the second part of my dissertation, I take advantage of complete genome sequence available for twelve species of the genus *Drosophila* to study patterns of gene family evolution and positive selection on protein coding genes in the *Drosophila* immune

system. I show that genes encoding proteins involved in recognition of invading microorganisms, specifically those involved in phagocytosis and cellular immunity, are much more likely to have experienced positive selection than either other immune system genes or non-immune genes. Furthermore, I show that rates of gene duplication and loss are extremely high among genes that encode the effector proteins that kill bacteria, suggesting flexibility in acquiring novel downstream components of the pathway during evolution, despite the very strong conservation of orthology among the signaling cascades that initiate the immune response. Taken together, the two parts of my dissertation demonstrate the importance of the structure of gene networks in evolution, and suggest a model where small differences in upstream components can have large phenotypic and evolutionary consequences, whereas even large differences in downstream components can be evolutionarily and phenotypically masked by redundancies among effectors.

BIOGRAPHICAL SKETCH

Timothy Sackton was raised in Lexington, MA, where he attended Lexington High School and first developed an interest in biology. After graduating high school in 1997, he attended Brown University. At Brown, he took an evolutionary biology class with Professor David Rand, which lead to several years of undergraduate research in the Rand lab, culminating in an undergraduate honor's thesis. Timothy Sackton graduated from Brown *magna cum laude* with a Bachelor's of Science degree in 2001. He then attended graduate school at Cornell, working with Professor Andrew Clark.

Dedicated to my wife, Kate, my muse and inspiration

ACKNOWLEDGMENTS

I would like to acknowledge a number of people, without whom this dissertation would not have been possible. My graduate research was funded by a Howard Hughes Medical Institute Predoctoral Fellowship, a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-0608196), as well as the National Institutes of Health grant R01 AI064950 to Andrew Clark and Brian Lazzaro. I had a number of collaborators on this work: Brian Lazzaro (Cornell), Todd Schlenke (Emory), Dan Hultmark (Umeå) and Jay Evans (USDA-ARS Beltsville) provided valuable contributions, especially to chapters 2 and 4. A large number of people helped collect the data presented in chapter 2: Anthony Fiumera, Todd Schlenke, Tracy Mak, Phil Olsen, Jenna Rose, Zoe Weiss, Mike Bosemeny, Cheryl Seidel, Bonnie Scurman, Xiaoyun Wang, Sean Hackett, and Yoko Hatta-Ohashi. Leah Greenspan, Bethany Schiller, and Jason Kwong were invaluable in helping to collect the data presented in chapter 3. Finally, Peter Schweitzer at the Cornell Life Sciences Core Laboratory helped with genotyping for chapter 2 and short-read sequencing for chapter 5. Computational resources provided by the Computational Biology Service Unit at the Cornell Theory Center were of critical help for a number of aspects of my dissertation work. I would like to thank Amanda Larracuenta, Nadia Singh, Anthony Fiumera, Todd Schlenke, Kurt McKean, Kristi Montooth, and all other members of the Clark lab, past and present, for providing an stimulating environment in which to conduct research. I would also like thank my committee, Charles Aquadro, Carlos Bustamante, and Martin Wiedmann, for invaluable advice and help along the way, and my advisor Andrew Clark for support, advice, and being the best mentor I could have hoped for. Lastly, I would like to thank my family, and especially my wife Kate Sackton, for being so supportive during my time in graduate school.

TABLE OF CONTENTS

Biographical Sketch.....	iii
Dedication.....	iv
Acknowledgements.....	v
List of Figures.....	vii
List of Tables.....	viii
Chapter 1.....	1
Chapter 2.....	18
Chapter 3.....	52
Chapter 4.....	82
Chapter 5.....	113

LIST OF FIGURES

Figure 1.1: Schematic of the <i>Drosophila</i> immune response.....	3
Figure 1.2: Phlyogeny of the sequenced <i>Drosophila</i> species.....	9
Figure 2.1: Mean bacterial load sustained after infection by one of four bacteria.....	21
Figure 2.2: Fraction of significant association tests by functional class.....	26
Figure 2.3: Associations at the GNBPI/GNBP2 locus.....	27
Figure 2.4: Expression of <i>pole hole</i> plotted against bacterial load.....	35
Figure 2.5: Induction of effector genes significantly correlates with bacterial load.....	38
Figure 3.1: Design of diallel crossing experiment.....	58
Figure 3.2: Additive effects attributable to selected markers after challenge.....	65
Figure 3.3: Regression of predicted against observed additive effects.....	67
Figure 3.4: Non-additive effects attributable to selected markers after challenge.....	65
Figure 3.5: Posterior mean bacterial load of genotypes from two diallel crosses.....	71
Figure 4.1: Variation in patterns of homology among immune system genes.....	85
Figure 4.2: Schematic of the <i>nimrod</i> and <i>eater</i> genes in 12 <i>Drosophila</i> species.....	88
Figure 4.3: Variation in positive selection among immune system genes.....	93
Figure 4.4: Positive selection in PRGP-LCa.....	97
Figure 4.5: Positive selection in the imd pathway.....	99
Figure 5.1: Number of reads mapped and filtered from cDNA sequencing data.....	117
Figure 5.2: Induction from cDNA sequencing compared to microarray data.....	120
Figure 5.3: Fraction of genes induced in microarray experiments that are induced based on cDNA sequencing.....	121
Figure 5.4: Relative expression of antimicrobial peptides between species.....	128
Figure 5.5: Lengths of expressed regions in <i>D. melanogaster</i> and <i>D. virilis</i>	135
Figure 5.6: Distribution of 3' UTR lengths in <i>D. melanogaster</i>	137

LIST OF TABLES

Table 2.1: Correlations of lines means between bacteria.....	22
Table 2.2: List of markers typed for association testing experiments.....	23
Table 2.3: Markers with significant associations with bacterial load.....	25
Table 3.1: Markers used in diallel cross experiments.....	56
Table 3.2: Estimated variance components from diallel crosses.....	63
Table 3.3: Parameters estimated from diallel simulations.....	73
Table 4.1: Distribution of positively selected sites among recognition proteins.....	96
Table 4.2: Distribution of positively selection sites among Relish and its interactors.....	101
Table 5.1: Binomial probability estimated for each induction state.....	119
Table 5.2: Number of expressed genes assigned to each induction state.....	119
Table 5.3: Induction of antimicrobial peptides in <i>D. melanogaster</i>	125
Table 5.4: Induction of antimicrobial peptides in <i>D. virilis</i>	126
Table 5.5: Genes associated with induced regions in <i>D. virilis</i> that lack homologs in <i>D. melanogaster</i>	130
Table 5.6: Association of induced expressed regions with gene models in <i>D. virilis</i>	138

CHAPTER 1

INTRODUCTION

Host-Pathogen Interactions and Their Evolutionary Consequences

Evolutionary biologists have long recognized that interactions between organisms have the potential to exert significant influence on evolutionary dynamics. In particular, competitive interactions between species imply that any relative advantage gained by one species will create selective pressures on interacting species. This dynamic underlies classic evolutionary theories, such as the Red Queen hypothesis (Van Valen 1973; Dawkins and Krebs 1979), which posit that rapid evolutionary change can result from these ‘arms race’ dynamics in competitive interactions between species.

Host-pathogen interactions are a particularly striking example of a competitive contest between species: at the most fundamental level pathogens compete with their hosts for control over the resources provided by the host organism. Just as these pathogenic organisms that are ubiquitous in nature, all organisms possess systems that serve to defend themselves against intruders. The ubiquity of host-pathogen interactions, and the generally severe consequences of infection on organismal fitness, imply that natural selection should be a major driver of patterns of genetic variation and evolutionary divergence at genes involved in host immunity. Indeed, many studies across a wide range of organisms have demonstrated that immune system genes evolve rapidly and adaptively at the protein level (*e.g.*, Hughes and Nei 1988; Murphy 1993; Schlenke and Begun 2003; Obbard *et al.* 2006; Tiffin and Moeller 2006; Sackton *et al.* 2007), suggesting a major role for natural selection in shaping patterns of polymorphism and divergence.

In the following chapters, I focus on two aspects of the genetics of host immune system evolution using *Drosophila* as a model organism. In Chapters 2 and 3, I focus on the genetic architecture of naturally occurring variation for immune function, in order to better understand why segregating genetic variation for this trait with presumably important consequences on organismal fitness persists in natural populations. In Chapters 4 and 5, I focus on the divergence of immune system genes in the genus *Drosophila*, both at the level of sequence divergence and at the level of divergence of gene content and copy number. While numerous studies have demonstrated that immune system genes are often rapidly evolving, the availability of comprehensive genomic resources in *Drosophila* allows a careful dissection of the role that network structure plays in shaping the evolution of the immune system.

The Drosophila Immune System

Over the last 15 years, a wealth of genetic and molecular studies have led to an increasingly clear understanding of the *Drosophila* immune system (reviewed in Lemaitre and Hoffmann 2007). This extensive knowledge, combined with unparalleled genomic resources, make *Drosophila* an ideal model system to study the evolution of complex networks, including immunity. Like all insects, *Drosophila* lack the antibody-mediated adaptive immunity that is the hallmark of vertebrates. Instead, they rely on a generalized set of innate host defenses (Figure 1.1): a humoral immune response that culminates in the production of antimicrobial peptides, and a cellular immune response mediated by circulating hemocytes that both phagocytose microorganisms and encapsulate larger foreign objects such as parasitoid eggs. The humoral response is initiated when circulating and membrane-bound recognition proteins bind to pathogen-associated molecular signatures, such as peptidoglycan from bacterial cell walls (Kaneko *et al.* 2004; Steiner 2004; Stenbak *et al.* 2004). These

recognition proteins include members of two gene families, the peptidoglycan recognition proteins (PGRPs, Werner *et al.* 2000; Choe *et al.* 2002; Gottar *et al.* 2002; Werner *et al.* 2003) and the Gram-negative binding proteins (GNBPs, Gobert *et al.* 2003; Wang *et al.* 2006). This bacterial recognition triggers a signaling cascade via

CIRCULATING HEMOCYTES

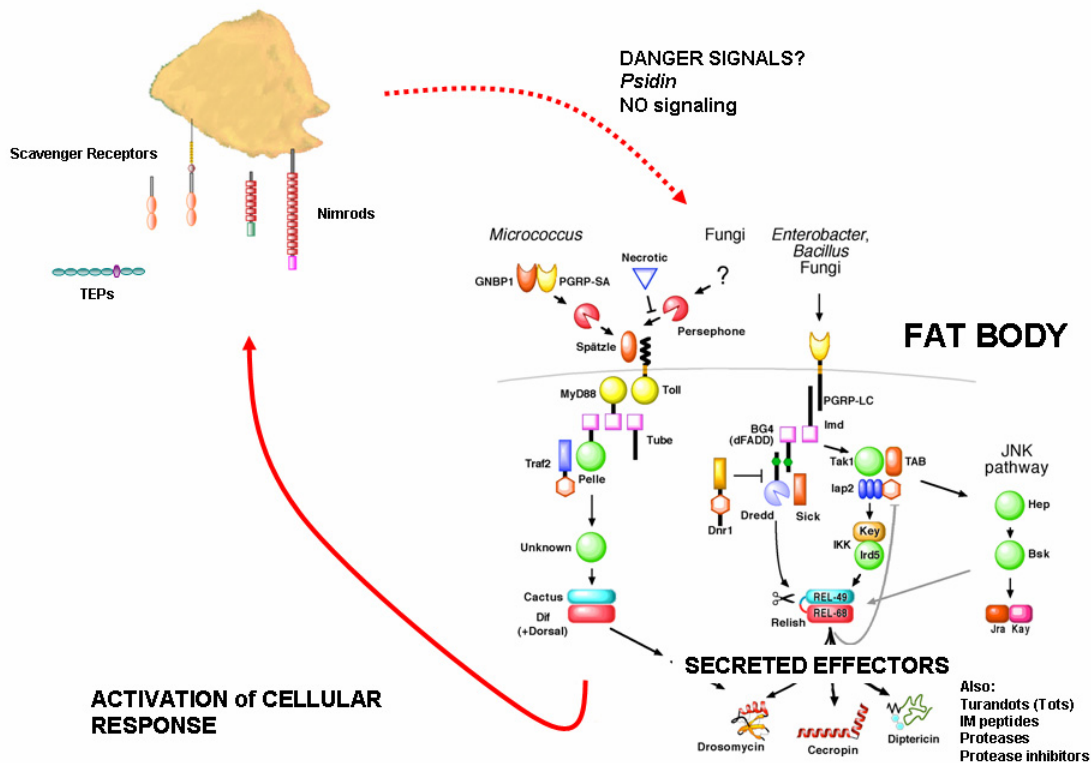


Figure 1.1. Schematic of the *Drosophila* immune response, showing major signaling pathways and constituent proteins. Modified from Hultmark (2003).

one of two primary immune signaling pathways, the Toll and Imd pathways (Lemaitre *et al.* 1995; Lemaitre *et al.* 1996; Boutros *et al.* 2002; De Gregorio *et al.* 2002a). Activation of these pathways, named for key constituent proteins, leads to the translocation of the NF-κB transcription factors Dorsal, DIF, and Relish into the nucleus (Dushay *et al.* 1996; Hedengren *et al.* 1999; Meng *et al.* 1999). These

transcription factors then up-regulate a suite of effector proteins, including most prominently antimicrobial peptides that have bactericidal activity (Imler and Bulet 2005), but also a number of other putative effectors such as the IM family of small peptides (Uttenweiler-Joseph *et al.* 1998), the Turandots (Ekengren and Hultmark 2001; Ekengren *et al.* 2001), iron binding proteins (Yoshiga *et al.* 1999; Dunkov and Georgieva 2006), and others (De Gregorio *et al.* 2001; Boutros *et al.* 2002; De Gregorio *et al.* 2002b).

The cellular response, and particularly phagocytosis, has received less attention than the humoral response, and only recently have the basic molecular mechanisms of *Drosophila* phagocytosis begun to be elucidated (see Stuart and Ezekowitz 2008 for a recent review). *Drosophila* possess three lineages of circulating hemocytes: two of them (crystal cells and lamellocytes) are primarily involved in melanization and encapsulation of parasitoid wasp eggs; the third, plasmatocytes, are a lineage of cells dedicated to a phagocytic role (Meister and Lagueux 2003; Meister 2004).

Plasmatocytes encode a number of surface receptors that mediate binding and engulfment of bacteria: the best characterized of these are the scavenger receptors (Pearson *et al.* 1995; Ramet *et al.* 2001) and the Nimrod family of EGF-like proteins (Kurucz *et al.* 2003; Kocks *et al.* 2005; Kurucz *et al.* 2007). Other molecules, such as the α -2-macroglobin-like Thioester containing proteins (TEPs), may play an important role as opsonins (Blandin and Levashina 2004; Blandin *et al.* 2004; Moita *et al.* 2005; Stroschein-Stevenson *et al.* 2006). Recent evidence has also suggested the importance of signaling between circulating hemocytes and the fat body (the main site of AMP production) in initiating a humoral response (Foley and O'Farrell 2003; Brennan *et al.* 2007).

Genetic architecture of immune function

It is now well established that natural populations of most organisms harbor substantial segregating genetic variation for immune function (*e.g.*, Carius *et al.* 2001; Niare *et al.* 2002; Lazzaro *et al.* 2004; Tinsley *et al.* 2006). In the face of pathogen-mediated selection pressure, the persistence of this variation requires explanation. One set of hypotheses involves the cost of immunity: if immune systems are costly to maintain and deploy, then overinvestment in immune function will be selectively disfavored. Ultimately, the evolutionarily optimal level of immunity need not be complete resistance. This hypothesis has a long history of theoretical support (Antonovics and Thrall 1994; Boots and Haraguchi 1999; Frank 2000; Rigby *et al.* 2002); experimental demonstration of costs of immunity also exist (*e.g.*, Webster and Woolhouse 1999; McKean *et al.* 2008), but the extent to which costs alone are sufficient to explain observed levels of variation is unclear. A related hypothesis proposes that there are trade-offs among resistance alleles, with some alleles increasing resistance to one pathogen at the cost of decreasing resistance to others. This kind of antagonistic pleiotropy can result in the selective maintenance of variation under certain conditions, but has been difficult to conclusively demonstrate for alleles mediating variation in immune function (but see Carius *et al.* 2001; Lazzaro *et al.* 2006).

In order to understand the maintenance of genetic variation, it is first necessary to understand the underlying genetic architecture of immunocompetence. Resistance to pathogens is a complex trait, mediated by many different subsystems (such as phagocytosis, melanization, and production of AMPs) that are interrelated in a complex regulatory network. Furthermore, immune traits are often condition dependent and show extensive genotype by environment interactions (*e.g.*, Mitchell *et*

al. 2005; Lazzaro *et al.* 2008). This poses a challenge in attempting to understand the basic underlying architecture of genetic variation in immunocompetence.

A fruitful first step has been to consider a simplified version of the problem, by using homozygous chromosome substitution lines and a candidate gene associating mapping approach to focus on underlying genetic variation attributable to putative immune genes on single chromosomes at a time in homozygous individual (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted). Using a set of second chromosome substitution lines, in which each line is homozygous for a different second chromosome isolated from nature, and all lines are isogenic for the other chromosomes, Lazzaro *et al.* (2004) mapped variation in bacterial load after infection with *Serratia marcescens* to candidate genes, primarily signaling and recognition genes. Using the same set of lines, Lazzaro *et al.* (2006) extended this work by examining statistical associations between genetic markers and bacterial load after infection with several additional bacterial species. This work revealed that the quantitative genetic basis of *D. melanogaster* antibacterial defense is complex and variable across infecting pathogens, suggesting that adaptive evolution in the *Drosophila* antibacterial immune system may be complicated by genotype x environment interactions and heterogeneity in prevalence of different pathogenic bacteria in time and space.

In Chapter 2, I describe work using a set of third chromosome substitution lines, in which we test for statistical associations between markers in candidate genes and variation in bacterial load after infection with each of four different bacteria. We show that markers in or near recognition genes are more likely to be statistically associated with differences in bacterial load, and provide support for previous observations that any effect of genetic variation in AMP is far too small to be observed in experiments such as these. Furthermore, we measure gene expression before and

after infection of over 300 known and candidate immune genes, and show that induction of genes encoding effector proteins significantly correlates with bacterial load. Taken together, these observations suggest a model where *trans* acting effects in upstream genes can be amplified by the pathway, leading to changes in the expression of large coregulated suites of downstream components and thus more significant associations with phenotype, whereas genetic variation in genes encoding proteins at the downstream terminus of the network, such as AMPs, has relatively little effect, as changes in any single effector protein appear to be of relatively small consequence individually.

While the strategy of measuring immune phenotypes in homozygous, inbred lines has advantages for association testing, it has important limitations as well. In the absence of information about the phenotype of heterozygotes, it is impossible to measure the relative importance of additive and dominance variation in contributing to differences in bacterial load phenotypes. In Chapter 3, I describe work that begins to address this limitation of previous work, using targeted diallel crosses to estimate parameters of a quantitative genetic model. We have constructed a series of targeted diallel crosses, using the previously characterized chromosome substitution lines as the source of the parental lines; for each diallel cross, we have genotypic information for candidate markers that associate with differences in bacterial load across the parental lines. From this set of crosses, we can estimate the additive and dominance contributions to bacterial load of particular targeted SNPs, by fitting a genetic model to the diallel cross phenotypes. These results underscore the complexity of the genetic architecture of the innate immune system in *Drosophila*.

Understanding the genetic basis of complex traits is a major research goal of a wide variety of fields, from animal breeding to human health (Lynch and Walsh 1998). In the context of the *Drosophila* immune system, the data presented in Chapters

2 and 3 underscores the importance of considering the complexity of the genetic architecture underlying complex traits. In Chapter 2, we show that the architecture of the genetic network has important consequences for mapping genotype to phenotype, suggesting that analysis of complex traits needs to take into account not just individual genes, but variation across whole pathways. Furthermore, in Chapters 2 and 3 we show that the genetic basis of variation in immune function in *Drosophila* is complex, with a wide range of additive and dominance effects. Furthermore, we document varying effects of single markers across different bacterial infections, suggesting that selective maintenance of variation in this trait may be possible.

Interspecific divergence of the immune gene regulatory

Despite the complexity of the genetic architecture of innate immune systems, and evidence that variation in immune system genes may be selectively maintained, numerous studies of genes involved in immune function across a wide variety of taxa have demonstrated that immune system genes remain one of the most rapidly evolving classes of genes between species (*e.g.*, Hughes and Nei 1988; Murphy 1993; Schlenke and Begun 2003; Obbard *et al.* 2006; Tiffin and Moeller 2006; Sackton *et al.* 2007). Given the strong selection pressure and evolutionary arms race dynamics that exist in host-pathogen relationships, this observation is not particularly surprising. However, much of the previous work has either focused on a few genes in limited species (*e.g.*, Hughes and Nei 1988; Begun and Whitley 2000), or on comparisons between very distant species with limited power to infer patterns of past selection on genes (Christophides *et al.* 2002; Evans *et al.* 2006; Waterhouse *et al.* 2007). This has led to the current situation, where a general consensus that immune system genes are rapidly evolving and a few dramatic case studies of particular genes (*i.e.*, MHC, Relish) have masked a lack of emphasis of the role of network structure in mediating how different

parts of the immune system respond differently to pathogen-mediated selection pressures.

In Chapter 4, I describe work based on the recent sequencing of ten additional species of *Drosophila* (Clark *et al.* 2007), bring the total number of sequenced

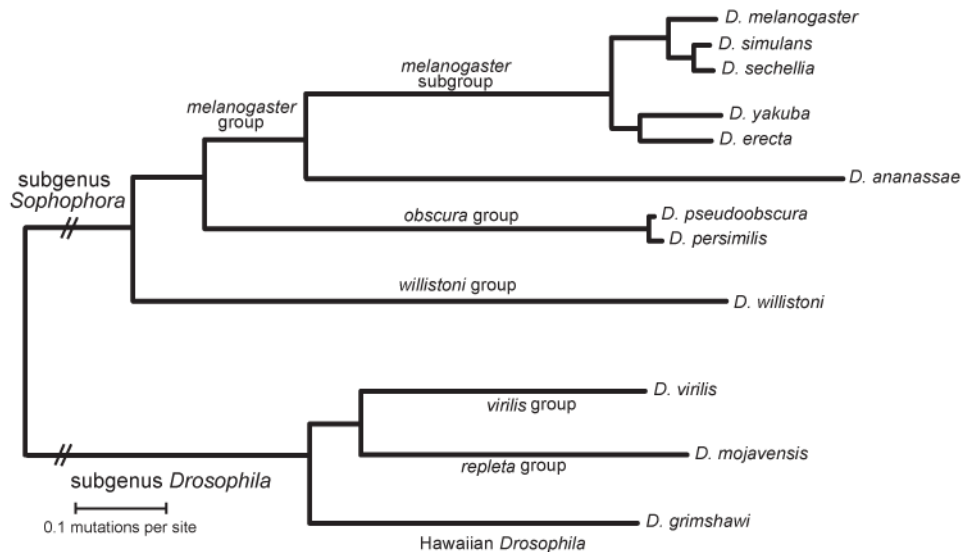


Figure 1.2. Phylogeny of the sequenced *Drosophila* species. The branch length between the two subgenera is not shown to scale. Data from Clark *et al.* (2007).

Drosophila genomes to twelve (Adams *et al.* 2000; Richards *et al.* 2005; Clark *et al.* 2007). With complete genome sequence from a large number of species that span a range of evolutionary distances (Figure 1.2), we were able to examine two aspects of evolutionary divergence across all genes currently known to have a role in the innate immune response. First, we used codon-based models of molecular evolution to compare rates of amino acid divergence and evidence for signatures of positive selection across different classes of immune system genes. We show that genes encoding proteins involved in pathogen recognition (specifically, those encoding putative phagocytic receptors that mediate recognition in the cellular immune

response) are much more likely to have experienced positive selection than genes encoding signaling or effector proteins. Second, we use birth-death models of gene family evolution to estimate rates of diversification and turnover of multigene families. Here, we find a different pattern: it is genes encoding effector proteins that show the most rapid rates of turnover and the highest extent of lineage-specific patterns of homology. This work thus suggests that network structure is a key determinant of patterns of interspecific divergence, particularly with regard to rates of gene duplication: effector proteins, at the bottom of the network, appear to be relatively unconstrained in terms of rates of gene duplication, whereas genes encoding members of signaling cascades show very low rates of duplication.

In order to understand the patterns of acquisition and loss of components of the immune system, it is necessary to move beyond homology-based annotation of immune components in non-model species, as by definition homology-based annotation cannot identify truly novel components of the innate immune system that have been recruited in one lineage. As a first step towards this goal, in Chapter 5 I describe work characterizing the transcriptional response to infection in *D. virilis* using short-read sequencing technology. By sequencing cDNA pools from infected and uninfected samples, we characterized genes that respond transcriptionally to infection, and identify components of the *D. virilis* immune system that appear to be evolutionary novelties restricted to *D. virilis* and its close relatives. We find evidence for the acquisition of several novel putative effectors, supporting the observations presented in Chapter 4 that suggest downstream pathway components are uniquely suited for rapid turnover, acquisition, and loss during evolution.

Conclusions

There exists a bulk of evidence showing that natural populations harbor substantial genetic variation for immune function, and that genes encoding immune system proteins often evolve by positive selection. However, a fuller picture of immune system evolution at the population and interspecific level requires integration of distinctions between genes with different roles in the innate immune network. In the following chapters, I hope to show that such an integrative approach can yield important insights into the role of selection in shaping patterns of polymorphism and divergence in the *Drosophila* innate immune system.

REFERENCES

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287(5461): 2185-2195.
- Antonovics J, Thrall PH (1994) Cost of Resistance and the Maintenance of Genetic-Polymorphism in Host-Pathogen Systems. *Proceedings of the Royal Society of London Series B-Biological Sciences* 257(1349): 105-110.
- Begun DJ, Whitley P (2000) Adaptive evolution of relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics* 154(3): 1231-1238.
- Blandin S, Levashina EA (2004) Thioester-containing proteins and insect immunity. *Molecular immunology* 40(12): 903-908.
- Blandin S, Shiao SH, Moita LF, Janse CJ, Waters AP *et al.* (2004) Complement-like protein TEPI is a determinant of vectorial capacity in the malaria vector *Anopheles gambiae*. *Cell* 116(5): 661-670.
- Boots M, Haraguchi Y (1999) The evolution of costly resistance in host-parasite systems. *Am Nat* 153(4): 359-370.
- Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Developmental Cell* 3(5): 711-722.
- Brennan CA, Delaney JR, Schneider DS, Anderson KV (2007) Psidin is required in *Drosophila* blood cells for both phagocytic degradation and immune activation of the fat body. *Current Biology* 17(1): 67-72.
- Carius HJ, Little TJ, Ebert D (2001) Genetic variation in a host-parasite association: Potential for coevolution and frequency-dependent selection. *Evolution* 55(6): 1136-1145.
- Choe KM, Werner T, Stoven S, Hultmark D, Anderson KV (2002) Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in *Drosophila*. *Science* 296(5566): 359-362.
- Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S *et al.* (2002) Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298(5591): 159-165.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.

- Dawkins R, Krebs JR (1979) Arms races between and within species. *Proceedings of the Royal Society of London Series B*. 205(1161): 489-511.
- De Gregorio E, Spellman PT, Rubin GM, Lemaitre B (2001) Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 98(22): 12590-12595.
- De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002a) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *The EMBO Journal* 21(11): 2568-2579.
- De Gregorio E, Han SJ, Lee WJ, Baek MJ, Osaki T *et al.* (2002b) An immune-responsive Serpin regulates the melanization cascade in *Drosophila*. *Developmental Cell* 3(4): 581-592.
- Dunkov B, Georgieva T (2006) Insect iron binding proteins: insights from the genomes. *Insect Biochemistry and Molecular Biology* 36(4): 300-309.
- Dushay MS, Asling B, Hultmark D (1996) Origins of immunity: Relish, a compound Rel-like gene in the antibacterial defense of *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 93(19): 10343-10347.
- Ekengren S, Hultmark D (2001) A family of Turandot-related genes in the humoral stress response of *Drosophila*. *Biochemical and Biophysical Research Communications* 284(4): 998-1003.
- Ekengren S, Tryselius Y, Dushay MS, Liu G, Steiner H *et al.* (2001) A humoral stress response in *Drosophila*. *Current Biology* 11(9): 714-718.
- Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL *et al.* (2006) Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology* 15(5): 645-656.
- Foley E, O'Farrell PH (2003) Nitric oxide contributes to induction of innate immune responses to gram-negative bacteria in *Drosophila*. *Genes & Development* 17(1): 115-125.
- Frank SA (2000) Specific and non-specific defense against parasitic attack. *J Theor Biol* 202(4): 283-304.

- Gobert V, Gottar M, Matskevich AA, Rutschmann S, Royet J *et al.* (2003) Dual activation of the *Drosophila* toll pathway by two pattern recognition receptors. *Science* 302(5653): 2126-2130.
- Gottar M, Gobert V, Michel T, Belvin M, Duyk G *et al.* (2002) The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* 416(6881): 640-644.
- Hedengren M, Asling B, Dushay MS, Ando I, Ekengren S *et al.* (1999) Relish, a central factor in the control of humoral but not cellular immunity in *Drosophila*. *Molecular Cell* 4(5): 827-837.
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335(6186): 167-170.
- Hultmark, D (2003). *Drosophila* immunity: paths and patterns. *Curr. Opin. Immunol.* 15: 12-19.
- Imler JL, Bulet P (2005) Antimicrobial peptides in *Drosophila*: structures, activities and gene regulation. *Chemical Immunology and Allergy* 86: 1-21.
- Kaneko T, Goldman WE, Mellroth P, Steiner H, Fukase K *et al.* (2004) Monomeric and polymeric gram-negative peptidoglycan but not purified LPS stimulate the *Drosophila* IMD pathway. *Immunity* 20(5): 637-649.
- Kocks C, Cho JH, Nehme N, Ulvila J, Pearson AM *et al.* (2005) Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in *Drosophila*. *Cell* 123(2): 335-346.
- Kurucz E, Zettervall CJ, Sinka R, Vilmos P, Pivarcsi A *et al.* (2003) Hemese, a hemocyte-specific transmembrane protein, affects the cellular immune response in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 100(5): 2622-2627.
- Kurucz E, Markus R, Zsamboki J, Folkl-Medzihradzky K, Darula Z *et al.* (2007) Nimrod, a putative phagocytosis receptor with EGF repeats in *Drosophila* plasmatocytes. *Current Biology* 17(7): 649-654.
- Lazzaro BP, Scurman BK, Clark AG (2004) Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. *Science* 303(5665): 1873-1876.
- Lazzaro BP, Sackton TB, Clark AG (2006) Genetic variation in *Drosophila melanogaster* resistance to infection: a comparison across bacteria. *Genetics* 174(3): 1539-1554.

- Lazzaro BP, Flores HA, Lorigan JG, Yourth CP (2008) Genotype-by-environment interactions and adaptation to local temperature affect immunity and fecundity in *Drosophila melanogaster*. PLoS Pathogens 4(3): e1000025.
- Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. Annu Rev Immunol 25: 697-743.
- Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA (1996) The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in *Drosophila* adults. Cell 86(6): 973-983.
- Lemaitre B, Kromer-Metzger E, Michaut L, Nicolas E, Meister M *et al.* (1995) A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the *Drosophila* host defense. Proceedings of the National Academy of Sciences of the United States of America 92(21): 9465-9469.
- Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits: Sinauer Associates. 980 p.
- McKean KA, Yourth CP, Lazzaro BP, Clark AG (2008) The evolutionary costs of immunological maintenance and deployment. BMC Evol Biol 8: 76.
- Meister M (2004) Blood cells of *Drosophila*: cell lineages and role in host defence. Current Opinion in Immunology 16(1): 10-15.
- Meister M, Lagueux M (2003) *Drosophila* blood cells. Cellular microbiology 5(9): 573-580.
- Meng XJ, Khanuja BS, Ip YT (1999) Toll receptor-mediated *Drosophila* immune response requires Dif, an NF-kappa B factor. Genes & Development 13(7): 792-797.
- Mitchell SE, Rogers ES, Little TJ, Read AF (2005) Host-parasite and genotype-by-environment interactions: Temperature modifies potential for selection by a sterilizing pathogen. Evolution 59(1): 70-80.
- Moita LF, Wang-Sattler R, Michel K, Zimmermann T, Blandin S *et al.* (2005) In vivo identification of novel regulators and conserved pathways of phagocytosis in *A. gambiae*. Immunity 23(1): 65-73.
- Murphy PM (1993) Molecular mimicry and the generation of host defense protein diversity. Cell 72(6): 823-826.

- Niare O, Markianos K, Volz J, Oduol F, Toure A *et al.* (2002) Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. *Science* 298(5591): 213-216.
- Obbard DJ, Jiggins FM, Halligan DL, Little TJ (2006) Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Current Biology* 16(6): 580-585.
- Pearson A, Lux A, Krieger M (1995) Expression cloning of dSR-CI, a class C macrophage-specific scavenger receptor from *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* 92(9): 4056-4060.
- Ramet M, Pearson A, Manfrulli P, Li X, Koziel H *et al.* (2001) *Drosophila* scavenger receptor CI is a pattern recognition receptor for bacteria. *Immunity* 15(6): 1027-1038.
- Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S *et al.* (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Research* 15(1): 1-18.
- Rigby MC, Hechinger RF, Stevens L (2002) Why should parasite resistance be costly? *Trends Parasitol* 18(3): 116-120.
- Sackton TB, Lazzaro BP, Clark AG (submitted) Genotype and gene expression associations with variation in the *Drosophila melanogaster* immune response. *PloS Biology*.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D *et al.* (2007) Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics* 39(12): 1461-1468.
- Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system evolution. *Genetics* 164(4): 1471-1480.
- Steiner H (2004) Peptidoglycan recognition proteins: on and off switches for innate immunity. *Immunological Reviews* 198: 83-96.
- Stenbak CR, Ryu JH, Leulier F, Pili-Floury S, Parquet C *et al.* (2004) Peptidoglycan molecular requirements allowing detection by the *Drosophila* immune deficiency pathway. *Journal of Immunology* 173(12): 7339-7348.
- Stroschein-Stevenson SL, Foley E, O'Farrell PH, Johnson AD (2006) Identification of *Drosophila* gene products required for phagocytosis of *Candida albicans*. *PLoS Biology* 4(1): e4.

- Stuart LM, Ezekowitz RA (2008) Phagocytosis and comparative innate immunity: learning on the fly. *Nat Rev Immunol* 8(2): 131-141.
- Tiffin P, Moeller DA (2006) Molecular evolution of plant immune system genes. *Trends in Genetics* 22(12): 662-670.
- Tinsley MC, Blanford S, Jiggins FM (2006) Genetic variation in *Drosophila melanogaster* pathogen susceptibility. *Parasitology* 132(Pt 6): 767-773.
- Uttenweiler-Joseph S, Moniatte M, Lagueux M, Van Dorsselaer A, Hoffmann JA *et al.* (1998) Differential display of peptides induced during the immune response of *Drosophila*: a matrix-assisted laser desorption ionization time-of-flight mass spectrometry study. *Proceedings of the National Academy of Sciences of the United States of America* 95(19): 11342-11347.
- Van Valen L (1973) A new evolutionary law. *Evolutionary Theory* 1(1): 1-30.
- Wang L, Weber AN, Atilano ML, Filipe SR, Gay NJ *et al.* (2006) Sensing of Gram-positive bacteria in *Drosophila*: GGBP1 is needed to process and present peptidoglycan to PGRP-SA. *The EMBO Journal* 25(20): 5005-5014.
- Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316(5832): 1738-1743.
- Webster JP, Woolhouse MEJ (1999) Cost of resistance: relationship between reduced fertility and increased resistance in a snail-schistosome host-parasite system. *Proceedings of the Royal Society of London Series B* 266(1417): 391-396.
- Werner T, Borge-Renberg K, Mellroth P, Steiner H, Hultmark D (2003) Functional diversity of the *Drosophila* PGRP-LC gene cluster in the response to lipopolysaccharide and peptidoglycan. *The Journal of Biological Chemistry* 278(29): 26319-26322.
- Werner T, Liu G, Kang D, Ekengren S, Steiner H *et al.* (2000) A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* 97(25): 13772-13777.
- Yoshiga T, Georgieva T, Dunkov BC, Harizanova N, Ralchev K *et al.* (1999) *Drosophila melanogaster* transferrin. Cloning, deduced protein sequence, expression during the life cycle, gene localization and up-regulation on bacterial infection. *European Journal of Biochemistry / FEBS* 260(2): 414-420.

CHAPTER 2

GENOTYPE AND GENE EXPRESSION ASSOCIATIONS WITH VARIATION IN THE *DROSOPHILA MELANOGASTER* IMMUNE RESPONSE

Introduction

Drosophila, like other insects, use a generalized immune response to combat pathogens. Unlike vertebrates, the insect immune response consists solely of an innate response, with cellular and humoral (cell-free) arms (reviewed in Lemaitre and Hoffmann 2007). Despite considerable knowledge of the molecular basis of the *Drosophila* immune response, and increasing understanding of the extent of natural genetic variation for immunocompetence in this system (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Tinsley *et al.* 2006), relatively little is known about the role of network structure in shaping the phenotypic consequences of genetic variation.

The *Drosophila* immune system provides an ideal opportunity to examine the consequences of genetic variation and differences among lines in patterns of gene expression in the context of a well-defined network. Linking genetic variation in transcriptional regulation to differences in complex phenotypes has the potential to illuminate mechanistic aspects of genotype-phenotype associations. Passador-Gurgel and coworkers (2007) identified several genes in which transcript levels significantly associate with survival times after exposure of *D. melanogaster* to nicotine. Other studies in *Drosophila* have identified transcriptional variation associated with male reproductive success (Drnevich *et al.* 2004), male body size (Coffman *et al.* 2005), aggressive behavior (Edwards *et al.* 2006) and locomotive behavior (Jordan *et al.* 2007). Nonetheless, linking genetic variation to phenotypic variation via

transcriptional changes has proven difficult (Fiumera *et al.* 2005; Harbison *et al.* 2005).

In *Drosophila*, the humoral response is initiated by the recognition of microbial cell wall component by proteins such as PGRPs and GNBP (Choe *et al.* 2002; Gobert *et al.* 2003; Werner *et al.* 2003). These proteins activate two primary signaling pathways, the Toll and Imd pathways. The Toll pathway is primarily activated after infection by fungi and Gram-positive bacteria, whereas the Imd pathway is primarily activated after infection by Gram-negative bacteria (Lemaitre *et al.* 1997; De Gregorio *et al.* 2002), although this specificity is not absolute (Hedengren-Olcott *et al.* 2004; Tanji *et al.* 2007). In addition to these primary signaling pathways, the JAK/STAT and JNK pathways are thought to play a role in immune response, largely as part of the general stress response and wound healing (Boutros *et al.* 2002; Agaisse and Perrimon 2004). Activation of the Toll and Imd signaling pathways leads to the translocation of NF- κ B transcription factors (Relish, DIF, Dorsal) to the nucleus where they drive transcription of effector genes such as those encoding antimicrobial peptides. These effectors are then released into the hemolymph, where they act to directly kill invading microorganisms (Imler and Bulet 2005).

Previously, we have examined associations between bacterial load after infection with each of four different bacteria and genetic markers (SNPs and indels) in candidate genes on the *Drosophila melanogaster* second chromosome (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006). Here, we extend that work to markers in candidate genes on the third chromosome. Furthermore, we examine both bacterial load and gene expression phenotypes, testing associations between genotype, sustained bacterial load, and transcription level of approximately 400 known and putative immune system genes.

Results

Genetic variation for immune function on the third chromosome in Drosophila

We examined a sample of 94 third-chromosome substitution lines for variation in bacterial load sustained 28 hours after infection with each of four different bacteria: two Gram-negative bacteria, *Serratia marcescens*, *Providencia rettgeri*; and two Gram-positive bacteria, *Enterococcus faecalis*, and *Lactococcus lactis* (Figure 2.1). In order to assess the effect of different third chromosomes on bacterial load phenotypes, we compared the likelihood of the data under a statistical model that allows for a main effect of genetic line to the likelihood of the data under a model that does not. Likelihood ratio tests reveal a large, highly significant effect of third chromosome line on phenotypic variation in bacterial load against all four bacteria (*S. marcescens*: $\chi^2 = 128.42$, d.f. = 1, $P < 2.2 \times 10^{-16}$; *P. rettgeri*: $\chi^2 = 263.88$, d.f. = 1, $P < 2.2 \times 10^{-16}$; *E. faecalis*: $\chi^2 = 51.533$, d.f. = 1, $P = 7.04 \times 10^{-13}$; *L. lactis*: $\chi^2 = 35.391$, d.f. = 1, $P = 2.70 \times 10^{-9}$). Genetic line explains 66.9% of the non-error variance (14.5% of the overall variance) for load sustained after *S. marcescens* infection and 58.3% (22.1%) for load sustained after *P. rettgeri* infection, but only 27.4% (7.2%) for *E. faecalis* and 19.5% (6.2%) for *L. lactis*; total variance in bacterial load is much higher for the two Gram-positive bacteria, with much of the increase being caused by larger among-replicate variance. The smaller fraction of variance attributable to line may simply be a consequence of this technical issue, relating to vagaries of the primary infections. The overall mean load sustained after infection also varies among bacteria, ranging from a low of 2,186 colony forming units (CFU) per fly 28 hours after infection with *S. marcescens* to a high of 653,436 CFU per fly after infection with *L. lactis*.

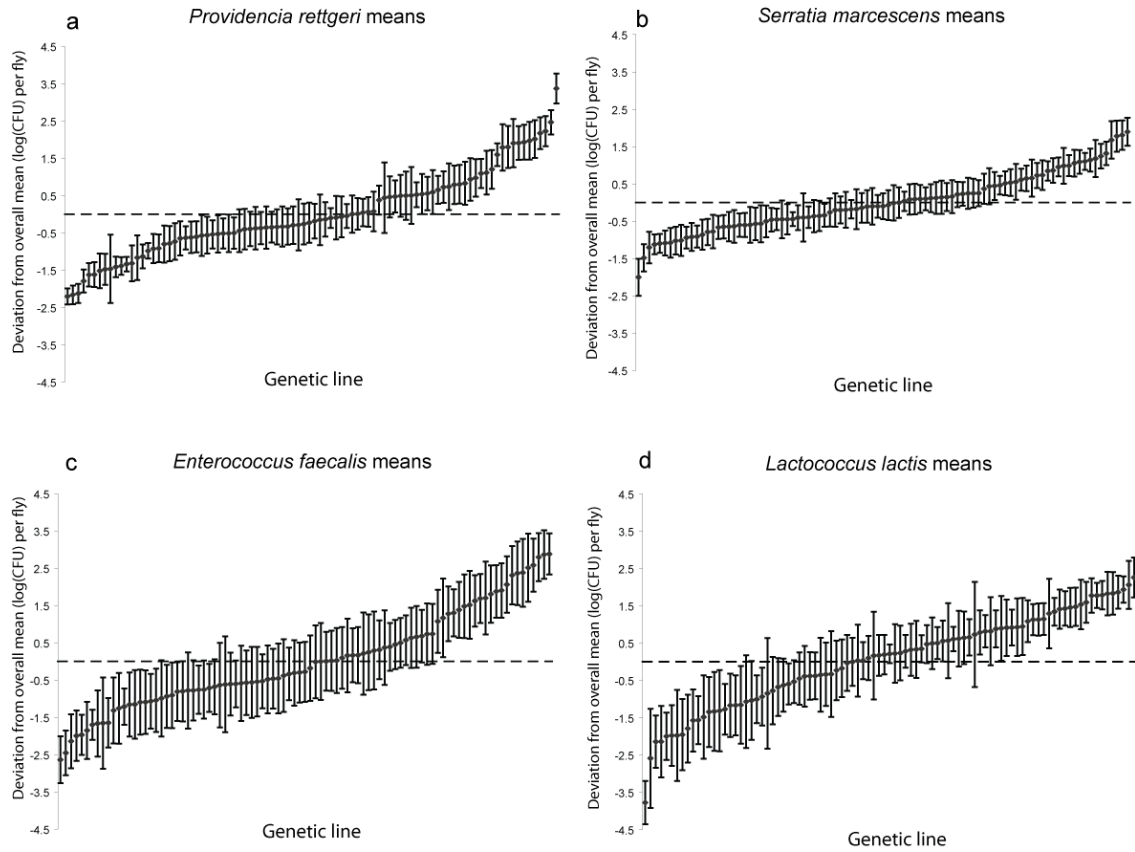


Figure 2.1: Mean bacterial load sustained 28 hours after infection by one of four different bacteria. Bacterial load is plotted as the deviation from the overall mean within each bacteria, adjusted for unbalanced data. The genetic lines are plotted independently in ascending rank order for each panel, and are not in the same order across panels.

Correlations of line means between bacteria (measured as Spearman's ρ) are generally moderate and positive (Table 2.1).

Table 2.1

Bottom diagonal: Spearman's ρ ; Top diagonal: P -value. Correlations were calculated using the function `cor.test` in R 2.6.0.

	<i>S. marcescens</i>	<i>P. rettgeri</i>	<i>L. lactis</i>	<i>E. faecalis</i>
<i>S. marcescens</i>		0.1369	0.02828	0.02527
<i>P. rettgeri</i>	+0.1561888		0.04021	0.00195
<i>L. lactis</i>	+0.2290430	+0.2145092		0.03012
<i>E. faecalis</i>	+0.2335280	+0.3202071	+0.2265	

Genotype-phenotype associations

We tested for statistical associations between bacterial load and genotypes at 137 polymorphisms in 25 genes and gene families on the third chromosome with known or suspected immune function. These included 6 antimicrobial peptide loci, 10 putative recognition loci (*GNBPs* and *PGRPs*), 8 known signaling loci, the Toll-like receptor Toll-9, and the iron-binding protein Transferrin 2 (Table 2.2). Our association test is based on mixed linear models: we assessed significance by comparing the observed regression coefficient for the marker in question to a null distribution generated from 1000 permuted data sets where the line phenotypes are randomly shuffled while preserving linkage disequilibrium among genetic markers. All statistical tests were implemented in R 2.6.0, as described in the Materials and Methods.

Across all bacteria, 42 tests (7.66%) are significant at a nominal α of 0.05, and 13 tests (2.37%) are significant at a nominal α of 0.01; in both cases, we observe a

significant excess of significant tests ($\alpha = 0.05$: $\chi^2 = 8.19$, d. f. = 1, P -value = 0.0042; $\alpha = 0.01$: $\chi^2 = 10.42$, d. f. = 1, P -value = 0.0012). Furthermore, 8 genotyped polymorphisms are significantly associated with bacterial load (at a nominal α of 0.05) after infection

Table 2.2

Functional Class	Locus	Cytological Position	Markers typed
Antimicrobial peptide	Attacin D	90B6	2
Antimicrobial peptide	CecAB	99E2	4
Antimicrobial peptide	CecC	99E2	3
Antimicrobial peptide	Drs	63D2	2
Antimicrobial peptide	DrsL	63D1--2	4
Antimicrobial peptide	dro2-5	63D1	3
Recognition	GNBP3	66E5	5
Recognition	GNBP1/GNBP 2	75D6	10
Recognition	PGRP-LA	67B1	7
Recognition	PGRP-LB	86E6	6
Recognition	PGRP-LC	67B1	10
Recognition	PGRP-LD	64E7--8	6
Recognition	PGRP-LF	67B1	4
Recognition	PGRP-SB1/2	73C1	11
Recognition	PGRP-SD	66A8	5
Signal Transduction	BG4	94A1	5
Signal Transduction	ECSIT	83C5	4
Signal Transduction	Rel	85C3	6
Signal Transduction	Toll	97D2	8
Signal Transduction	ird5	89B1	5
Signal Transduction	pll	97E11	5
Signal Transduction	spz	97E1	7
Signal Transduction	tub	82A5	2
Iron binding	Tsf2	69C4--5	4
Toll-like	Toll-9	77B6	9

with different bacteria in independent experiments. Three of these markers (GNBP_1041, GNBP_3696, GNBP_3768), all in either *GNBP1* or *GNBP2*, are significantly associated with *P. rettgeri* load at a 5% FDR. All 15 SNPs associated with load sustained after infection by one bacterium at a nominal α of 0.01, or after infection with two distinct bacteria each at a nominal α of 0.05, are presented as Table 2.3.

Significant tests (nominal $\alpha = 0.01$) are not randomly distributed among bacteria: 84.6% of the significant cases represent associations between genotype and bacterial load after infection with Gram-negative bacteria (*S. marcescens* and *P. rettgeri*). This excess could be driven by a biological difference in the response of *D. melanogaster* to Gram-positive and Gram-negative bacteria that results in less variation among lines in load after infection with Gram-positive bacteria. A more likely explanation, though, is that we lack power to detect significant associations against Gram-positive bacterial load due to higher among-replicate variance.

Significant associations (nominal $\alpha = 0.01$) are also not evenly distributed within functional classes of the immune system: all of the 15 polymorphisms associated with bacterial load listed in Table 2.3 are found within recognition or signaling genes. Across all polymorphisms, there are significant differences in the fraction of significant associations at a nominal α of 0.05 among functional classes (Figure 2.2; $\chi^2 = 11.76$, d. f. = 2, $P = 0.0028$). Polymorphisms in recognition genes are most commonly associated with differences in bacterial load (11.72% of tests), followed by polymorphisms in signaling genes (4.9% of tests). Polymorphisms in effector genes are very rarely associated with differences in bacterial load (only 2.27% of tests). This result remains significant if a nominal α of 0.01 is used to determine significance of individual associations ($\chi^2 = 6.51$, d. f. = 2, $P = 0.039$). Furthermore,

Table 2.3:

Minor allele frequency, effect size, and p-values for the 15 SNPs with either a significant association at $P < 0.01$ against one bacteria, or $P < 0.05$ against two bacteria. Significant associations are in bold. Further details on each SNP are described in Sackton *et al.*, submitted.

Marker	Major Allele	Minor Allele	MAF	<i>E. faecalis</i>		<i>L. lactis</i>		<i>P. rettgeri</i>		<i>S. marcescens</i>	
				Effect size (minor-major)	Perm. P	Effect size (minor-major)	Perm. P	Effect size (minor-major)	Perm. P	Effect size (minor-major)	Perm. P
GGBP75D_1041	C	A	0.3095	-0.1602	0.577	0.2366	0.343	0.8500	0.001	0.4238	0.019
GGBP75D_1209	G	A	0.3614	-0.1244	0.653	0.2417	0.300	0.7451	0.010	0.3165	0.067
GGBP75D_3350	A	G	0.4878	-0.0967	0.727	-0.0755	0.757	-0.6213	0.021	-0.3430	0.050
GGBP75D_3386	C	A	0.4405	-0.0695	0.807	0.0952	0.670	0.7543	0.004	0.2598	0.118
GGBP75D_3696	G	A	0.4000	0.0179	0.943	-0.0114	0.968	0.7505	0.001	0.4925	0.006
GGBP75D_3768	A	T	0.4405	0.1348	0.650	0.0816	0.727	0.7737	<0.0001	0.3509	0.030
GGBP75D_4856	C	G	0.3855	-0.4208	0.121	0.0412	0.855	0.7866	0.004	-0.0075	0.967
PGRPLC_5624	A	T	0.0556	-0.1324	0.822	-0.9413	0.050	-0.4711	0.400	-0.9638	0.010
PGRPLC_6635	A	C	0.2333	-0.0088	0.977	-0.6806	0.006	-0.3701	0.223	-0.4875	0.010
PGRPLC_884	T	C	0.4368	-0.5146	0.050	-0.2977	0.193	-0.2909	0.263	-0.3414	0.034
PGRPLF_2711	A	G	0.2184	0.1554	0.663	0.6085	0.020	0.2422	0.417	0.4967	0.024
pil_1214	A	T	0.0778	-0.1118	0.808	0.1722	0.680	0.0785	0.857	0.7359	0.008
pil_2099	T	C	0.3133	0.1218	0.688	-0.1504	0.549	-0.2457	0.406	-0.4793	0.006
Rel_793	T	C	0.1818	0.8160	0.011	0.1204	0.682	0.0466	0.902	0.1826	0.401
spz_595	G	T	0.1264	1.0754	0.005	0.3097	0.340	0.2731	0.488	0.1681	0.496

this rank order is maintained if we rank functional classes by the fraction of genes with at least one significant test (instead of the overall fraction of significant tests within each functional class), or if the functionally less-well-understood genes *Tsf2* and *Toll-9* are excluded.

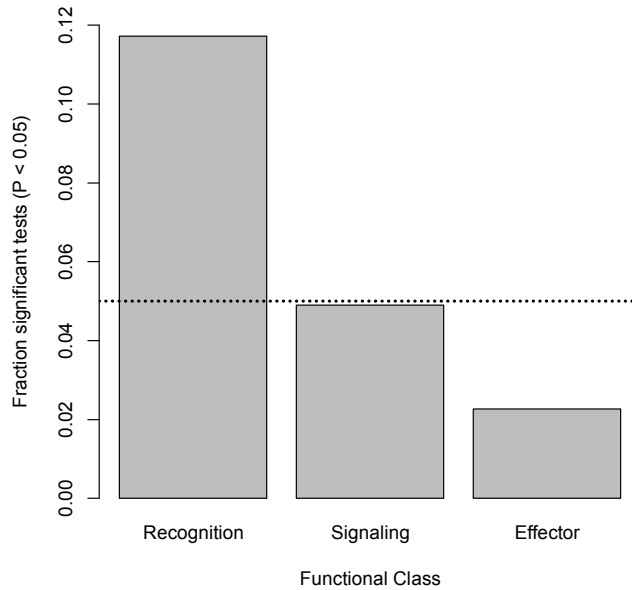


Figure 2.2:

Fraction of association tests within each functional class (pooled across all bacteria) that are significant at a nominal α of 0.05 (dotted line). The difference among classes is significant ($\chi^2 = 11.76$, d. f. = 2, $P = 0.0028$)

Polymorphism at the GNB75D locus, consisting of the genes *GNBP1* and *GNBP2*, is particularly striking in the extent and significance of associations with resistance to Gram-negative bacteria (Figure 2.3). Seven of the 10 SNPs at this locus are nominally significantly associated with variation in bacterial load after infection with *P. rettgeri*, although average linkage disequilibrium is high at this locus (average pairwise $r^2 = 0.303$; average pairwise $D' = 0.636$). Four of those seven SNPs are also significantly

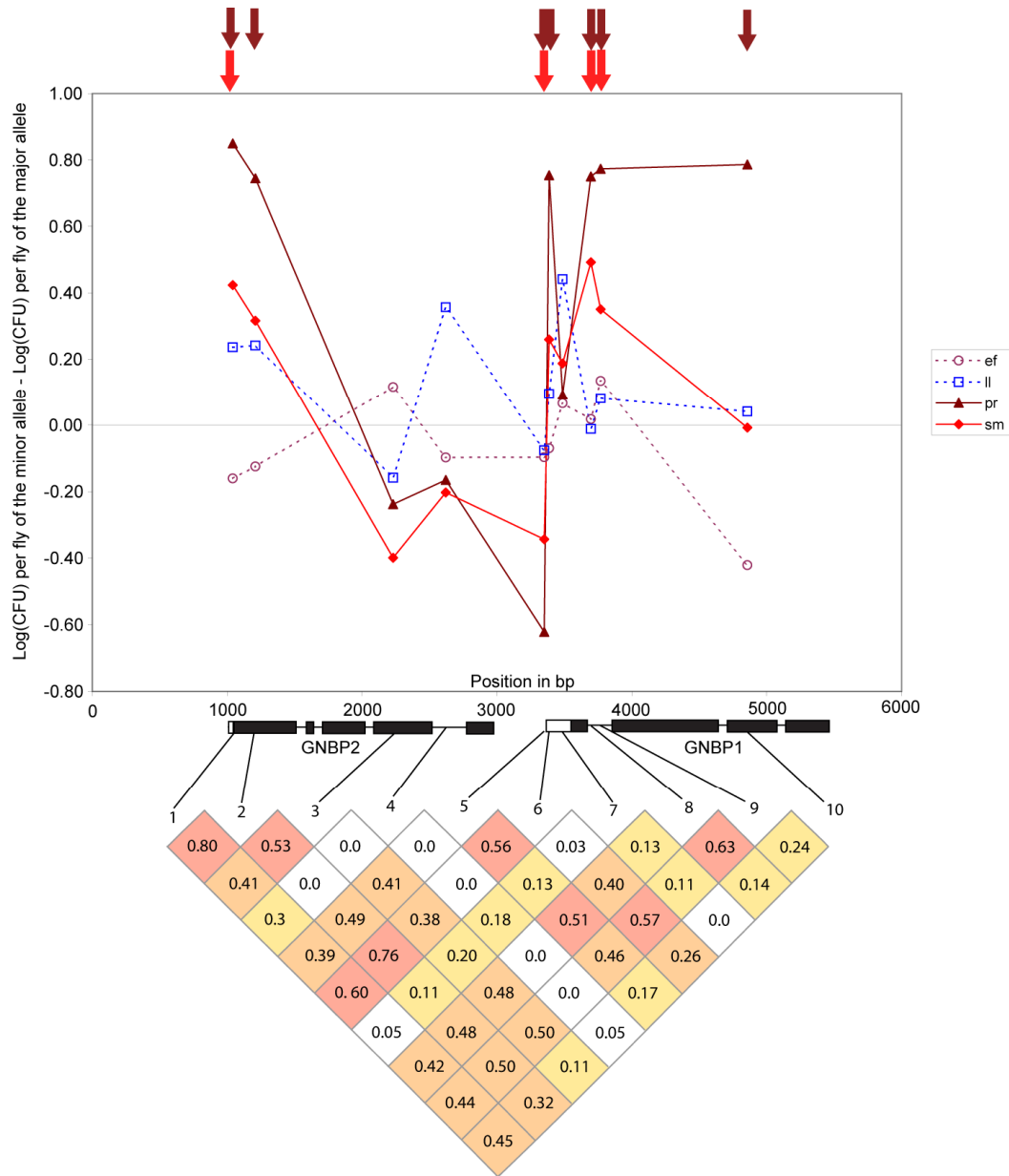


Figure 2.3

Upper panel: Plot of the effect size for each of the 10 SNPs genotyped at the GNB locus in chromosomal band 75D. Gram-negative bacteria are shown in dashed lines, Gram-positive as solid lines. Arrows above the main graph indicate significance at a nominal α of 0.05. Lower panel: Pattern of linkage disequilibrium among the 10 genotyped SNPs. Grid shows r^2 values, shaded by value: >0.50 red, 0.25-0.50 orange, 0.10-0.25 yellow.

associated with differences in bacterial load after infection with *S. marcescens*. These include one SNP in the 3' UTR of *GNBP2* (GNBP75D_1041), one SNP in the 5' UTR of *GNBP1* (GNBP75D_3350), and a pair of SNPs in the first intron of *GNBP1* (GNBP75D_3696 and GNBP75D_3768).

The haplotype structure at the GNBP75D locus is unusual for *D. melanogaster*. Despite spanning more than 2 kb, the four SNPs mentioned previously are found in only 6 of the 16 possible haplotypes in 91 of the 94 genetic lines (the remaining three lines have unique haplotypes). There are two major haplotypes (A-A-A-T and C-G-G-T) at frequencies of 0.244 and 0.449 respectively. When the phenotypes of the lines that carry these two haplotypes are compared directly using nonparametric tests, the A-A-A-T haplotype has a significantly higher median bacterial load after infection with either *S. marcescens* (median_{AAAT} = 13.16, median_{CGGT} = 12.26, Mann-Whitney U $P = 0.0002993$) or *P. rettgeri* (median_{AAAT} = 8.12, median_{CGGT} = 7.48, Mann-Whitney U $P = 0.036$).

PGRP-LC is another recognition gene with repeatable evidence for a significant association between SNPs and bacterial load. In this case, genotypes at three out of 10 SNPs are associated with variation in bacterial load against at least two different bacteria. A SNP marker approximately 125 bp upstream of the transcriptional start site of *PGRP-LC* is associated with resistance against both *E. faecalis* and *S. marcescens*. A SNP marker in the third exon of splice variant *PGRP-LC-RB* (intronic in splice variants *PGRP-LC-RA* and *PGRP-LC-RC*) is associated with variation in bacterial load against *L. lactis* and *S. marcescens*, with the same allele associated with lower bacterial load against both bacteria. Another SNP marker in the fourth exon (in the PGRP domain) of *PGRP-LC-RA* (in the intron of *PGRP-LC-RA* and *PGRP-LC-RC*) is also associated with variation in bacterial load against *L. lactis* and *S. marcescens*. These two SNP markers are in linkage disequilibrium ($r^2 = 0.193$, $P =$

7.92×10^{-4}), but neither is in significant linkage disequilibrium with the upstream marker.

Marker by sex interactions

Empirical and theoretical work (Zuk 1990; Rolff 2002; McKean and Nunney 2005) suggest that immune function may differ between the sexes, as males and females make different resource allocation decisions between immune defense and reproductive output. These observations lead to the hypothesis that the genetic basis of the immune response may depend on sex: indeed, these kinds of genotype by sex interactions have been observed for other quantitative traits in *D. melanogaster* (e.g., Mackay 2004). To test this hypothesis, we compared the likelihood of our observed bacterial load data under a model with a Sex by Marker interaction to the likelihood of the data under a model without such an effect (but retaining the main effects of Sex and Marker). To assess the significance of the resulting likelihood ratio test statistics, we used a null distribution of likelihood ratio test statistics calculated by permuting the data 1000 times.

We find little evidence for significant effects of marker by sex interactions on bacterial load. While 6.93% of tests are significant at a nominal α of 0.05, a weakly significant excess over the null expectation ($\chi^2 = 4.32$, d. f. = 1, $P = 0.0377$), only 0.91% of tests are significant at a nominal α of 0.01, which is not different from the null expectation ($\chi^2 = 0.042$, d. f. = 1, $P = 0.8367$). While it is possible that there are weak marker by sex interactions that we do not have the power to detect in this experiment, we believe that such effects are likely to be small compared to main effects of SNP across sexes. There is a clear effect of sex: males have consistently lower bacterial loads irrespective of genotype, consistent with the hypothesis that the sex bias in immune function is phenotypically plastic in *Drosophila*, and depends on food and mate availability (McKean and Nunney 2005).

Measuring gene expression using BeadChips

In order to understand the mechanistic basis of differences in immune phenotypes linked to genetic variation on the third chromosome, we measured gene expression of approximately 700 transcripts in males from a subset of 30 of the 94 phenotyped chromosome 3 substitution lines. Using custom-designed Illumina BeadChip microarrays, we measured transcription under three different conditions (uninfected, 8 hours post *S. marcescens* infection [Sm-infected], and 8 hours post *E. faecalis* infection [Ef-infected], where *S. marcescens* and *E. faecalis* were chosen arbitrarily to represent Gram-negative and Gram-positive bacteria respectively). We selected the subset of lines to assay to be biased towards the tails of the phenotypic distribution in order to enhance our power to detect correlations between transcript abundance and phenotype. We normalized and log-transformed expression values as described in the methods. For most analyses, we focused on the Ef-induced (Ef-infected minus uninfected) and Sm-induced (Sm-infected minus uninfected) samples.

In addition to quantifying the 329 genes with a known or putative immune function (including 172 genes with some characterized function and 157 genes predicted to have a role in immunity based on transcriptional induction after infection), the BeadChip microarrays include genes involved in metabolism (139) and sex/reproduction (164), as well as 69 probesets consisting of housekeeping gene controls, and genes involved in insecticide resistance. Full details of the BeadChip design are described in the Materials and Methods. For most analyses, we focus on the 329 immune genes on the BeadChips, although in some cases we use the other genes as controls.

Genotype-expression associations

Although with only 30 lines we have limited power to detect significant associations between SNPs and gene expression variation, we tested for significant

associations by comparing a mixed model with a fixed effect of SNP to one with just a fixed intercept. Because permutations are not computationally feasible for the large number of tests required for this analysis, we assessed significance by comparing the likelihood ratio test statistic to a standard χ^2 distribution. Overall, 3.55% and 2.99% of genotype-expression association tests are significant at a nominal α of 0.01 in the Sm-induced and Ef-induced samples, respectively. In both cases it is highly improbable to obtain this many significant tests purely by chance (Sm-induced: $\chi^2 = 1414$, d. f. = 1, $P < 2.2 \times 10^{-16}$; Ef-induced: $\chi^2 = 972$, d. f. = 1, $P < 2.2 \times 10^{-16}$). The same pattern holds if we consider the absolute expression level in the Ef-infected, Sm-infected, and uninfected samples individually (data not shown).

Because we assumed the null distribution of the test statistic follows an asymptotic chi-square distribution, it is possible that the excess of significant P -values we observe is primarily due to the mis-specification of the null distribution. We expect that polymorphisms in genes known to have a role in the immune system will be more likely to affect expression of immune-related genes than expression of other genes on the BeadChip. Indeed, for the Ef-induced sample, we see significantly more significant tests among immune-related genes than other genes (0.0325 vs. 0.0274; $\chi^2 = 21.6874$, d.f. = 1, $P = 3.206 \times 10^{-6}$), although this is not the case for the Sm-induced sample (but note that “non-immune” genes may still be responding transcriptionally to infection). Thus, while it appears that some of the genotyped SNPs in this study have significant effects on gene expression, particularly for the Ef-induced sample, limiting our experiment to 30 lines reduces our power to detect significant associations. Nonetheless, there are 304 and 350 associations between genotypes and induction of immune genes after *E. faecalis* and *S. marcescens* infection respectively significant at a 10% false-discovery-rate. However, given the uncertainty in the true estimates of

significance, we focus in the next two sections on overall patterns of genotype-expression associations.

Stronger genotype-expression associations in cis than in trans

In order to understand how genetic variation for transcriptional differences fits into the structure of the innate immune network, we first considered the extent that expression-genotype associations are more likely to be in close proximity to the SNP marker in question (*cis*) than distant from the marker (*trans*). For the Ef-induced sample, there is a significant excess of nominally significant tests between gene expression and *cis* markers compared to *trans* markers (Fisher's Exact Test $P = 0.0278$, Odds Ratio = 1.99). For the Sm-induced sample, the trend is in the same direction, although it is not significant (Fisher's Exact Test $P = 0.0857$, Odds Ratio = 1.69). However, in both cases there are still many more *trans* effects than expected by chance (Ef-induced: $\chi^2 = 2257$, $P < 2.2 \times 10^{-16}$; Sm-induced: $\chi^2 = 2803$, $P < 2.2 \times 10^{-16}$), suggesting propagation of expression differences throughout the signaling network.

Significant associations tend to follow the predicted network structure

Because a considerable amount is known about the transcriptional feedback relationships in innate immune networks, we can make some predictions about the expected direction of associations between genotypes and variation in gene expression of specific genes that go beyond simple binary classifications of *cis* vs. *trans* effects. Most generally, we expect that markers in signaling genes should predict expression of effector genes much more often than markers in effector genes associate with expression of signaling genes. In other words, we believe that genetic differences in signaling genes could lead to differential expression of effector genes, but that genetic differences in effector genes do not result in feedback that influences transcription of signaling genes. For both the Ef-induced and Sm-induced samples, we see exactly this

pattern (Ef-induced: Fisher's Exact Test $P = 0.000102$, Odds Ratio = 1.939; Sm-induced: Fisher's Exact Test $P = 0.00558$, Odds Ratio = 1.625).

We further dissected the role of network structure in the pattern of associations between gene expression and SNPs by examining the number of significant associations between markers in either the Toll or Imd pathway and expression of genes in other signaling pathways. On the BeadChips, we have representatives from the Toll, Imd, JAK/STAT, JNK, Ras, p38, and Notch signaling pathways. We compared the observed number of tests significant at $\alpha = 0.01$ to the expected number based on chance alone, using χ^2 tests. For the Ef-induced sample, we observe a significant excess (over chance expectations) of associations between markers in genes in the Toll pathway and induction of signaling genes in the Toll pathway ($P = 1.32 \times 10^{-13}$) and the JAK/STAT pathway ($P = 3.05 \times 10^{-14}$); we also observe an excess of significant associations between markers in genes in the Imd pathway and induction of signaling genes in the Imd pathway ($P = 0.00159$) and the Toll pathway ($P = 0.0292$, non-significant after Bonferroni correction).

For the Sm-induced sample, we see a similar pattern. There is a significant excess of significant associations between markers in Toll pathway genes and induction of genes in the Toll pathway ($P = 1.32 \times 10^{-13}$), and to a lesser extent induction of genes in the Imd pathway ($P = 0.0341$) and the JAK/STAT pathway (0.0496), although the latter two P -values do not survive a Bonferroni correction. Markers in Imd pathway genes are significantly more likely than expected by chance to be associated with induction of genes in the Imd pathway ($P = 0.0219$) and the JAK/STAT pathway ($P = 0.00102$) after infection with *S. marcescens*.

Because the numbers of markers in signaling genes represent a relatively limited sample, some caution should be used in interpreting these results. Nonetheless, these data suggest that, in addition to self-regulation of both the Toll and Imd

signaling pathways by components of the pathway, there is some cross-talk between the Toll, Imd, and JAK/STAT pathways, although there seems to be relatively little cross-talk between either of the Toll or Imd pathways and the JNK pathway, at least at the time point we examined (8 hours after infection). Given genetic variation for flux through the pathway, these patterns of autoregulation and cross-regulation may have the effect of amplifying the phenotypic consequences of minor genetic variants.

Quantitative trait transcripts

Considerable recent interest has focused on identifying not just genetic markers that associate with quantitative variation in phenotypes, but also transcripts where abundance correlates with phenotypes of interest (Fiumera *et al.* 2005; Passador-Gurgel *et al.* 2007; Emilsson *et al.* 2008). These attempts have had mixed success, with some studies failing to find any significant correlations between transcript abundance and phenotype (*e.g.*, Fiumera *et al.* 2005) and others finding some evidence for significant associations (*e.g.*, Passador-Gurgel *et al.* 2007).

Here, we used a simple regression of the induction of immune-related transcripts against either *E. faecalis* bacterial load (for Ef-induced sample) or *S. marcescens* bacterial load (for Sm-induced sample) to attempt to detect expression-phenotype associations. In this analysis, induction correlates with bacterial load for very few transcripts. Only the inductions of *Attacin C* and *Drosocin* after *E. faecalis* infection correlate with *E. faecalis* bacterial load at a false discovery rate of 10%. Induction levels after *S. marcescens* infection do not appear to correlate with *S. marcescens* load for any transcripts, although uninfected expression level of *pole hole* (D-Raf) associates with *S. marcescens* load at a FDR of 0.0035, the most significant transcriptional association in our dataset (Figure 2.4). Interestingly, *pole hole* is necessary for the proliferation or survival of circulating hemocytes in *D. melanogaster*

(Luo *et al.* 2002; Evans *et al.* 2003) suggesting that flies with lower levels of *phl* transcription may have fewer hemocytes, and be less able to resist infection.

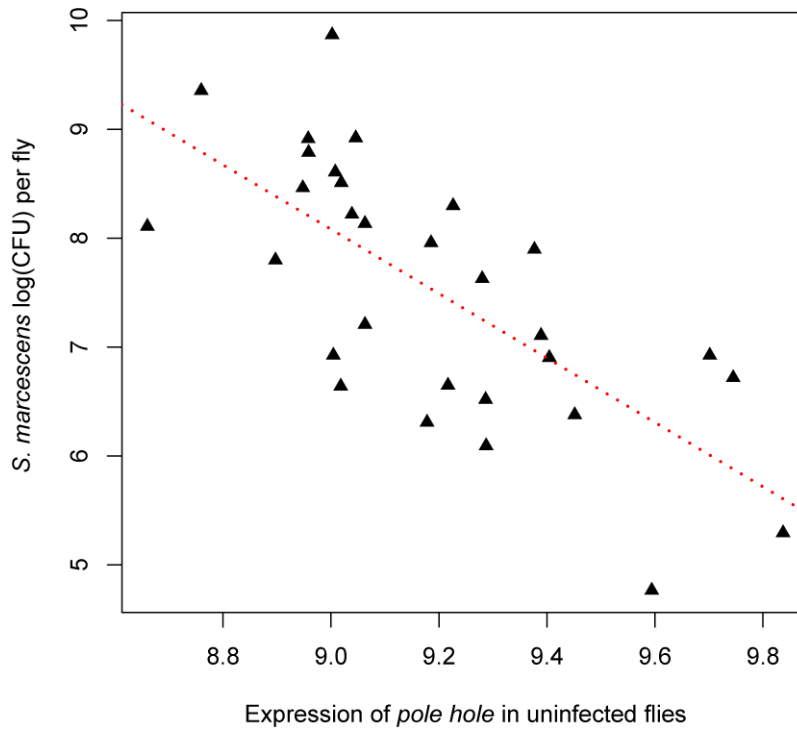


Figure 2.4:

Normalized expression of *pole hole* in uninfected flies (see materials and methods) plotted against bacterial load 28 hours after infection with *S. marcescens*.

Nonetheless, this analysis suggests that expression level of individual genes, measured as either induction after infection or as absolute expression in uninfected flies, is not a good predictor of bacterial load phenotypes. However, given the structure of the immune network, this observation may not be surprising. The immune system is a highly coregulated system, in which small changes in expression of upstream components can be amplified among downstream genes. Furthermore, correlated transcription of many effectors could indicate that what is biologically

relevant is not the transcription of individual genes, but rather the overall extent to which the immune system, in whole or in part, is transcriptionally activated after infection. In order to test this hypothesis, we considered whether principal components obtained from the correlation matrix among transcriptional profiles of subsets of genes predict phenotype. As an added advantage, the method of principal components reduces the dimensionality of large datasets, improving power.

Principal component analysis

Our initial hypothesis is that the most important transcriptional determinant of phenotype is the extent to which effector proteins are induced after infection. To measure this, we initially constructed a set of principal components (PCs) from the 61 genes in our dataset with a known or putative “effector” function. These include antimicrobial peptides, components of the phenoloxidase cascade, lysozymes, putative iron-sequestration proteins, and some less-well-characterized genes such as the Turandots. For both the Sm-induced dataset and the Ef-induced dataset, the variance explained by the first principal component is substantially higher than the variance explained by any other, and so we have focused on the first PC when looking for correlations with phenotypes.

This first PC estimated from the effector genes in the Ef-induced sample is significantly positively correlated with *E. faecalis* bacterial load (Figure 2.5A; $\beta = 74.8$, $F_{1,28} = 7.309$, $P = 0.01153$), explaining just over 20% of the variance among lines in resistance to *E. faecalis* ($r^2 = 0.207$). This PC is dominated by negative loadings of several antimicrobial peptide genes (*Mtk*, *DptB*, *AttC*, *Drs*) and genes encoding several uncharacterized peptides known to be induced by infection (*IM23*, *IM10*, *TotM*, *IM2*, *IM4*, *IMI*). Thus, this analysis suggests that genetic lines that induce antimicrobial peptides (and potentially related peptides) more strongly (*i.e.*,

have a lower PC1) sustain a lower bacterial load and thus are better able to resist infection.

We also examined the Sm-induced sample using a similar procedure. However, we do not see any correlation between the first PC from the effector genes in the Sm-induced sample and *S. marcescens* bacterial load ($\beta = 7.819$, $F_{1,28} = 0.2491$, $P = 0.6216$), despite the fact that the Sm-induced PC is quite similar to the Ef-induced PC in terms of loadings. *Serratia marcescens* is resistant to the antimicrobial effects of Cecropins (Samakovlis *et al.* 1990), Drosocins, and Defensins (Nehme *et al.* 2007) suggesting that this bacterium may be particularly resistant to *Drosophila* antimicrobial defenses and providing a plausible hypothesis for the lack of effect of variation in effector gene induction on variation in bacterial load. The resistance of *S. marcescens* to antimicrobial peptides may also explain the disproportionate effect of expression level of the hematopoietic gene *pole hole* on resistance to *S. marcescens* infections, as cellular immunity may be the main mechanism of *D. melanogaster* resistance to *S. marcescens*.

Genetic associations with PC1

A major challenge of quantitative genetics in *Drosophila* has been to link genetic polymorphisms to phenotypes via differences in expression. In this study, we have shown correlations between transcript abundance and phenotype, as well as correlations between genotype and phenotype. To look for genotype-expression-phenotype correlations, we focused on the *E. faecalis* bacterial load phenotype and the Ef-induced expression sample, and asked whether any of the SNPs that have nominally significant correlations with bacterial load are also correlated with the effector induction PC1. Of the eight SNPs with at least nominal associations between genotype and phenotype ($P < 0.05$), we find that one of them, PGRPSD_494, is also statistically associated with effector induction PC1 (Figure 2.5B; $\beta = 0.0235$, $F_{1,27} =$

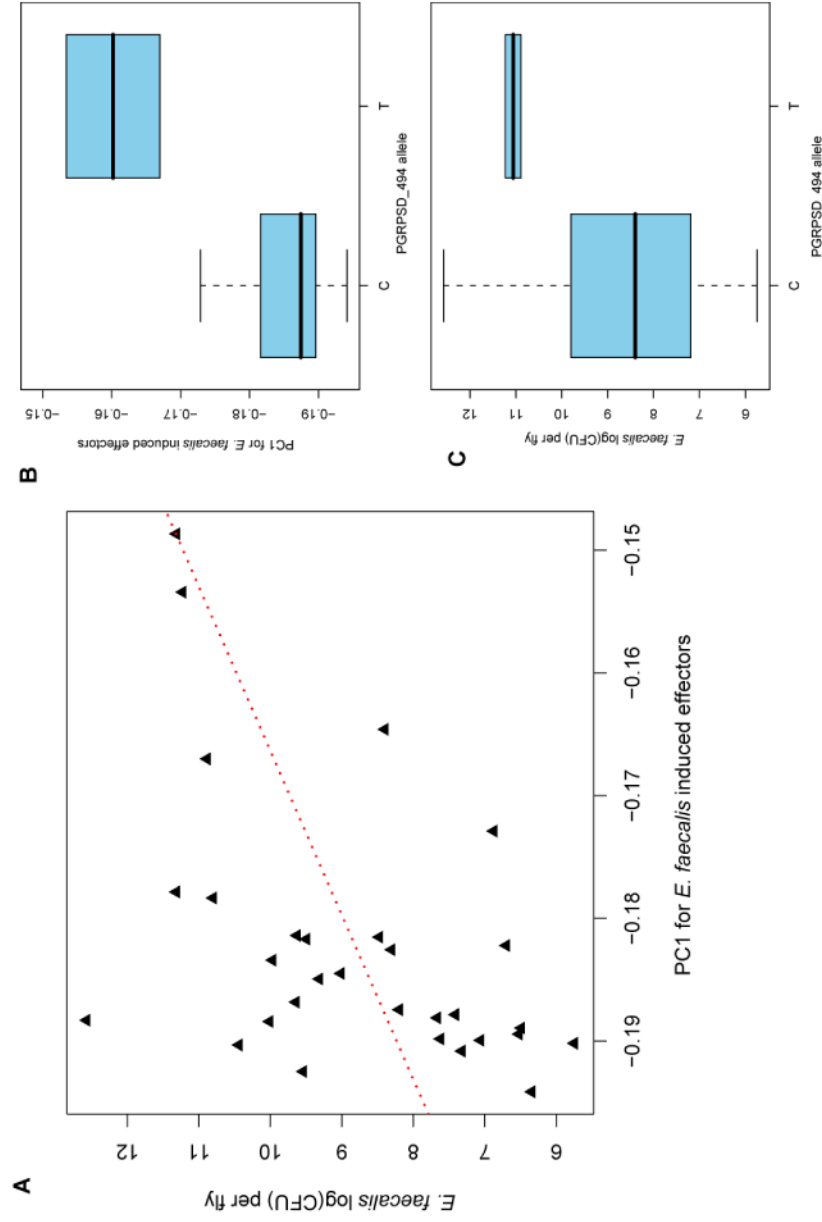


Figure 2.5:

A) First principal component from the expression of effectors in the Ef-induced sample, plotted against bacterial load 28 hours after *E. faecalis* infection. **B)** Box plot of the first principle component from the expression of effectors in the Ef-induced sample for the two allelic states at the PGRP-SD_494 marker. **C)** Box plot of bacterial load 28 hours after *E. faecalis* infection for the two allelic states at the PGRP-SD_494 marker.

11.4, $P = 0.002237$), explaining nearly 30% of the variance in this principal component ($r^2 = 0.297$).

The PGRPSD_494 marker is a C/T polymorphism located approximately 500 bp upstream of the transcriptional start site of *PGRP-SD*. The T allele is associated with both a higher bacterial load after infection (Ef load_{T-C} = 0.6741; $P = 0.02$) and with lower induction of antimicrobial peptides (higher PC1; Figure 2.5C). *PGRP-SD* has been shown to have a role in the recognition of some Gram-positive bacteria, including *E. faecalis* (Bischoff *et al.* 2004), and our data suggest that naturally occurring variation in *PGRP-SD* may in fact mediate the strength of the transcriptional response to infection, and thus the ability of the fly to resist infection. While this site does not appear to be significantly associated with induction or naïve expression of *PGRP-SD* in our data, it is surprisingly associated, at a 10% FDR, with induction levels of 73 of the 329 immune genes we assayed; no other SNP in our dataset is associated with induction levels of more than 14 genes, and most are associated with induction levels of fewer than 10 genes. We suggest that small differences in the transcriptional dynamics between the two alleles of *PGRP-SD* (smaller than what we can detect in our sample) can cause large impacts on the downstream transcriptional dynamics and organismal phenotype, due to the amplification effect of signaling cascades.

Discussion

The pursuit of an understanding of underlying determinants of phenotypic variation has a long history in *Drosophila* and other organisms (Lynch and Walsh 1998). More recently, the availability of high-throughput gene expression microarrays has generated interest in correlating variation in transcript abundance across genetic lines with differences in phenotypes (Drnevich *et al.* 2004; Coffman *et al.* 2005; Fiumera *et al.* 2005; Harbison *et al.* 2005; Passador-Gurgel *et al.* 2007). However,

datasets that include both genotype information and transcriptional variation have been rare (but see Fiumera *et al.* 2005; Wentzell *et al.* 2007). In this paper, we have focused on attempting to predict immunocompetence in *D. melanogaster* from SNPs in candidate genes and transcript abundance, guided by the known structure of the innate immune network (Lemaitre and Hoffmann 2007).

Although some of the associations we find are under test conditions with relatively high false-discovery rates, we can take advantage of the replication of our experiment across four different bacterial strains to increase confidence in our associations. We identify two loci in particular that appear to harbor genetic variation that is repeatably associated with differences in bacterial load phenotypes, both encoding proteins involved in bacterial recognition.

One of these loci contains the closely linked genes *GNBP1* and *GNBP2*. Several SNPs representing a single major haplotype are associated with differences in bacterial load after infection with both *S. marcescens* and *P. rettgeri*. It is somewhat unexpected to suggest a role for variation at *GNBP1* in resistance against Gram-negative bacteria, as *GNBP1* has only been shown to be involved recognizing Gram-positive bacteria and activating the Toll signaling pathway (Pili-Floury *et al.* 2004; Wang *et al.* 2006). However, the major haplotype spans both genes, making it impossible to determine the causal variant, and no definitive role for *GNBP2* is known. The second of these loci encodes *PGRP-LC*. *PGRP-LC* is the major receptor in the Imd pathway (Choe *et al.* 2002; Gottar *et al.* 2002; R.,met *et al.* 2002; Choe *et al.* 2005). Three SNPs at this locus are associated with differences in bacterial load after infection with *S. marcescens*, *E. faecalis*, and *L. lactis*. The observation that variation in the Imd pathway, canonically thought to be principally involved in resistance to Gram-negative bacteria, appears to associate with differences in bacterial load after infection with Gram-positive bacteria suggests that the innate immune

network is dynamic with extensive feedback and cross-talk. This pattern is further demonstrated by the pattern of associations between genotype and gene expression: there are significantly more associations than expected by chance between SNPs in both the Toll and Imd pathways and signaling genes outside those pathways (particular in the JAK/STAT pathway).

Most of the significant associations between SNPs and phenotype that we observe in this study are in genes encoding recognition proteins, primarily *PGRPs* and *GNBPs*. This study, combined with previous association studies between immunocompetence and polymorphisms on the second chromosome (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006), allows us to infer general patterns about the genetic architecture of immunocompetence in *Drosophila*. The most striking observation is the near-complete lack of significant associations, even without correcting for multiple tests, in antimicrobial peptides. In this study and in the previous studies, we genotyped 204 markers covering every known antimicrobial peptide in *D. melanogaster*. Only a single marker (CecC_1660), a noncoding SNP downstream of *CecC*, has a nominal *P*-value less than 0.05, and even that marker is extremely unlikely to be a true association, as the association neither survives multiple test correction nor is observed in multiple experiments. Taken together, these studies provide convincing evidence that any effect of genetic variation in *D. melanogaster* AMP genes is far too small to be observed in experiments such as these. This observation supports the previous inference from genetic evidence that *Drosophila* AMPs are at least partially functionally redundant (Tzou *et al.* 2002).

A different picture is painted when considering the effect of variation across lines in overall transcript abundance. Here, the total induction of effector genes (primarily AMPs and other induced peptides such as the Turandots) appears to correlate with bacterial load, at least after *E. faecalis* infection. The combination of

genetic polymorphism, bacterial load phenotypes, and transcript abundance thus allows us to propose a model of the genetic architecture of immunocompetence informed by the structure of the innate immune network. Genetic variation in genes encoding proteins at the top of the network (such as recognition proteins) can be amplified by the pathway (as demonstrated by the association between variation at *PGRP-SD* and the Ef-induced effector PC1), leading to more significant associations with phenotype. However, genetic variation in genes encoding proteins at the bottom of the network, such as AMPs, has relatively little effect, as changes in any single effector protein do not seem to cause large enough effects on phenotype to be detectable in experiments of the scale we have performed. Since there appears to be relatively little feedback between SNPs in effector proteins and transcription of upstream genes (as demonstrated by the dearth of associations between effector SNPs and signaling gene transcripts), these SNPs probably have relatively little impact *trans*. Taken together, this model predicts that differences among lines in upstream proteins will likely have larger fitness consequences in terms of the response to infection, and that single mutations in antimicrobial peptides are likely to be of relatively little consequence. By combining expression data, genetic data, and knowledge of network structure, we can gain a much better understanding of the phenotypic consequences of genetic variation than any one component could provide alone.

Materials and Methods

Drosophila lines and bacterial stocks

We evaluated ninety-four lines of *D. melanogaster* for resistance to infection against each of four different bacteria. These lines are originally derived from a natural collection of wild-caught *D. melanogaster* from State College, PA by Anthony Fiumera. Each line in the panel is homozygous for an individual third chromosome

isolated from the natural population and substituted into a common genetic background. The construction of these lines is described in more detail in Fiumera *et al.* (2007).

The *D. melanogaster* lines in this study were challenged with each of four different bacteria, two Gram-positive and two Gram-negative. The Gram-positive bacteria used are the *E. faecalis* and *L. lactis* strains described in Lazzaro *et al.* (2006). The Gram-negative bacteria used are the *S. marcescens* strain described in Lazzaro *et al.* (2006), and *Providencia rettgeri* described in Juneja and Lazzaro (accepted pending revision).

Survey sequencing and genotyping

We ascertained markers to be typed by sequencing the complete coding region and 1-2 kb upstream of 25 candidate loci (listed in Table 2.2) from 8 lines. We assembled sequencing reads into contigs using Sequencher and manually identified SNPs and indels to assay in the full panel of 94 lines. We used three different methods for genotyping our panel of lines. Approximately half of the markers were genotyped using SNPlex (Applied Biosystems, Foster City, CA) and the remaining markers were genotyped using pyrosequencing assays, SNPstream (Beckman Coulter, Fullerton, CA), or fRLFP (Lazzaro *et al.* 2002). A small number of markers were genotyped with both SNPlex and pyrosequencing; for the rare cases where the genotype call disagreed, we used the SNPlex call. After genotyping, SNPs were filtered to produce a set of 137 usable markers (136 SNPs and 1 indel): markers with a minor allele frequency < 0.05 were dropped, and only one marker (chosen at random) was kept from any pair with LD (measured by r^2) > 0.90 . Annotation information for each SNP, including the genotyping method used to assay each SNP in the 94 lines, as well as genotype calls for each line and linkage disequilibrium between each pair of genotypes are described in Sackton *et al.* (submitted).

Bacterial infections

We infected the 94 *D. melanogaster* lines in a complete-block design, with each line infected on each of three different days. On each day, each line was infected by one of 3 to 5 infectors at random, and a different infector infected each line on each day. Typically 2-3 replicates per line per sex were obtained on each day, for a total of 12-18 replicate data points for each *D. melanogaster* line. The entire experiment was repeated independently for each bacterial challenge. Flies were artificially infected by septic pinprick as described previously (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006). Briefly, we pierced the thoraces of individual *D. melanogaster* aged 3–5 days post-eclosion with a 0.1-mm dissecting pin (Fine Science Tools, Foster City, CA) coated in liquid culture ($OD_{600} = 1.0 \pm 0.2$) of the bacterium of interest, delivering an average of 4×10^3 bacteria to each fly (not shown). *Drosophila* were maintained at 22°–24°C on a rich dextrose medium for the duration of the experiment. To measure bacterial load, we homogenized same-sex trios of flies 28 hours post-infection in 500 μ L of sterile LB and then quantitatively plated the homogenates on standard LB agar plays using robotic spiral platers manufactured by Spiral Biotech (Bethesda, MD) and Don Whitley Scientific (Fredrick, MD). We incubated the plates overnight at 37°C and then estimated the concentration of viable bacteria in each homogenate using the colony counting systems associated with each plater. Prior to plating, we diluted homogenates of *L. lactis* 1000-fold, homogenates of *P. rettgeri* 100-fold, and homogenates of *E. faecalis* 10-fold, all in sterile LB, in order to correct for anticipated high bacterial loads. Our estimates of bacterial load per fly were transformed to correct for these dilutions before analysis. Mean bacterial load sustained by each line against each of the four bacteria is presented in Figure 2.1.

BeadChip design

We selected 329 immune genes for inclusion on the custom Illumina BeadChips based on a number of criteria, including evidence for transcriptional regulation by infection in previous microarray experiments, genetic or molecular evidence for a role in immunity, and homology to known immune proteins in *D. melanogaster* or other organisms. The remaining 384 non-immune genes were selected either as controls or for other experimental reasons. Each gene is represented by two different probes, each of which is represented by an average of 30 beads on the array, giving an extremely high degree of technical replication.

Expression infections

We selected a total of 30 lines for our expression analysis, biased towards the upper and lower tails of the phenotypic distribution. Males of each line were either infected with *S. marcescens* with *E. faecalis*, as described above, or left uninfected, and then frozen 8 hours after treatment. We extracted total RNA using Trizol (Invitrogen Corp., Carlsbad, CA) following standard protocols, then made cDNA and amplified RNA samples following the BeadChip protocol.

BeadChip hybridizations and data normalization

RNA samples were hybridized to BeadChips following standard protocols and scanned. After scanning, we normalized the data using the qspline method in the beadarray package for R. Mean probability of detection and signal intensity of control genes were used as hybridization quality control: for samples that failed to pass quality control checks, cDNA synthesis, RNA amplification, and hybridization were repeated from the original RNA extractions. Normalized induction after *E. faecalis* and *S. marcescens* infection, as well as unnormalized expression data from all treatments (Ef-infected, Sm-infected, Naïve) are presented as Dataset S2 and Dataset S3, respectively.

Statistical analysis

In order to test for associations between genotype and phenotype, we analyzed the following model using the package lme4 in R 2.6.0,

$$Y_{ijklmn} = \text{Sex}_i + \text{Allele}_j + \text{Line}_k + \text{Day}_l + \text{Infector}_m + \text{Plater}_n + \varepsilon \quad (1)$$

where Y is bacterial load, Sex ($i=1,2$) and Allele ($j=1,2$) are main effects, and Line ($k=1,94$), Day ($l=1,3$), Infector ($m=1,5$), and Plater ($n=1,2$) are random effects. To assess significance, we compared the regression coefficient for the Allele term to the null distribution obtained by permuting the genotype vector assigned to each line 1000 times and reanalyzing the data with the same model.

In order to test for sex*marker interactions, we used a similar approach. In this case, we compared the likelihood of the data under the null model specified by equation (1) to likelihood of the data under the following alternative model:

$$Y_{ijklmn} = \text{Sex}_i + \text{Allele}_j + (\text{Sex} * \text{Allele})_{ij} + \text{Line}_k + \text{Day}_l + \text{Infector}_m + \text{Plater}_n + \varepsilon \quad (2)$$

where all terms are as described above. To assess significance, we compared the likelihood ratio test statistic obtaining by comparing the null and alternative models to the empirical null distribution of likelihood ratio test statistics obtained by analyzing 1000 permuted datasets in which the genotype vector assigned to each line was shuffled.

To test for associations between genotype and expression, we compared the likelihood of the data under the following linear model:

$$Y_{ij} = \text{Allele}_i + \text{Probe}_j + \varepsilon \quad (3)$$

where Y is the normalized induction of a given gene, Probe ($j=1,2$) is a random effect representing the two probes on the array for each gene, and Allele ($i=1,2$) is the fixed main effect of interest, to the likelihood of the data under the null model that retains the random effect of Probe but includes only a fixed intercept. As the number of tests is far too large for permutations to be computationally feasible, we used the `anova()` function in `lme4` to assess the significance of the alternative model using a likelihood ratio test.

In order to test for correlations between transcript abundance and phenotype, we used two approaches. In the first approach we tested each transcript against phenotype individually, using a simple linear regression (with the model `Load = Expression`) and assessing significance assuming the standard null distribution for the F statistic. In the second approach, we generated principal components from *a priori* subsets of transcripts, using the `prcomp()` function in R, and then assessed the correlation between the first principle component and bacterial load using a simple linear regression.

To correct for multiple testing, we used an false-discovery-rate (FDR) approach, as described in the Results section, implemented using the `p.adjust()` function in R.

REFERENCES

- Agaisse H, Perrimon N (2004) The roles of JAK/STAT signaling in *Drosophila* immune responses. *Immunological Reviews* 198: 72-82.
- Bischoff V, Vignal Cc, Boneca IG, Michel T, Hoffmann JA *et al.* (2004) Function of the *Drosophila* pattern-recognition receptor PGRP-SD in the detection of Gram-positive bacteria. *Nature Immunology* 5(11): 1175-1180.
- Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Developmental Cell* 3(5): 711-722.
- Choe K-M, Lee H, Anderson KV (2005) *Drosophila* peptidoglycan recognition protein LC (PGRP-LC) acts as a signal-transducing innate immune receptor. *Proceedings of the National Academy of Sciences of the United States of America* 102(4): 1122-1126.
- Choe K-M, Werner T, Stöven S, Hultmark D, Anderson KV (2002) Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in *Drosophila*. *Science* 296(5566): 359-362.
- Coffman CJ, Wayne ML, Nuzhdin SV, Higgins LA, McIntyre LM (2005) Identification of co-regulated transcripts affecting male body size in *Drosophila*. *Genome Biology* 6(6): R53-R53.
- De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *The EMBO Journal* 21(11): 2568-2579.
- Drnevich JM, Reedy MM, Ruedi EA, Rodriguez-Zas S, Hughes KA (2004) Quantitative evolutionary genomics: differential gene expression and male reproductive success in *Drosophila melanogaster*. *Proceedings Biological Sciences* 271(1554): 2267-2273.
- Edwards AC, Rollmann SM, Morgan TJ, Mackay TFC (2006) Quantitative genomics of aggressive behavior in *Drosophila melanogaster*. *PLoS Genetics* 2(9): e154-e154.
- Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature* 452(7186): 423-428.

- Evans CJ, Hartenstein V, Banerjee U (2003) Thicker than blood: conserved mechanisms in *Drosophila* and vertebrate hematopoiesis. *Developmental Cell* 5(5): 673-690.
- Fiumera AC, Dumont BL, Clark AG (2005) Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. *Genetics* 169(1): 243-257.
- Fiumera AC, Dumont BL, Clark AG (2007) Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. *Genetics* 176(2): 1245-1260.
- Gobert V, Gottar M, Matskevich AA, Rutschmann S, Royet J *et al.* (2003) Dual activation of the *Drosophila* toll pathway by two pattern recognition receptors. *Science* 302(5653): 2126-2130.
- Gottar M, Gobert V, Michel T, Belvin M, Duyk G *et al.* (2002) The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* 416(6881): 640-644.
- Harbison ST, Chang S, Kamdar KP, Mackay TFC (2005) Quantitative genomics of starvation stress resistance in *Drosophila*. *Genome Biology* 6(4): R36-R36.
- Hedengren-Olcott M, Olcott MC, Mooney DT, Ekengren S, Geller BL *et al.* (2004) Differential activation of the NF-kappaB-like factors Relish and Dif in *Drosophila melanogaster* by fungi and Gram-positive bacteria. *The Journal of Biological Chemistry* 279(20): 21121-21127.
- Imler J-L, Bulet P (2005) Antimicrobial peptides in *Drosophila*: structures, activities and gene regulation. *Chemical Immunology and Allergy* 86: 1-21.
- Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TFC (2007) Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biology* 8(8): R172-R172.
- Juneja P, Lazzaro BP (accepted pending revision) Novel species of *Providencia* isolated as infections of wild *Drosophila melanogaster*. *Intl J Syst Evol Microbiol.*
- Lazzaro BP, Scurman BK, Clark AG (2004) Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. *Science* 303(5665): 1873-1876.
- Lazzaro BP, Sackton TB, Clark AG (2006) Genetic variation in *Drosophila melanogaster* resistance to infection: a comparison across bacteria. *Genetics* 174(3): 1539-1554.

- Lazzaro BP, Scurman BK, Carney SL, Clark AG (2002) fRFLP and fAFLP: medium-throughput genotyping by fluorescently post-labeling restriction digestion. *Biotechniques* 33(3): 539-540, 542, 545-536.
- Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annual Review of Immunology* 25: 697-743.
- Lemaitre B, Reichhart JM, Hoffmann JA (1997) *Drosophila* host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 94(26): 14614-14619.
- Luo H, Rose PE, Roberts TM, Dearolf CR (2002) The Hopscotch Jak kinase requires the Raf pathway to promote blood cell activation and differentiation in *Drosophila*. *Molecular Genetics and Genomics* 267(1): 57-63.
- Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*: Sinauer Associates.
- Mackay TFC (2004) The genetic architecture of quantitative traits: lessons from *Drosophila*. *Current Opinion in Genetics & Development* 14(3): 253-257.
- McKean KA, Nunney L (2005) Bateman's principle and immunity: phenotypically plastic reproductive strategies predict changes in immunological sex differences. *Evolution* 59(7): 1510-1517.
- Nehme NT, Li,geois S, Kele B, Giammarinaro P, Pradel E *et al.* (2007) A model of bacterial intestinal infections in *Drosophila melanogaster*. *PLoS Pathogens* 3(11): e173-e173.
- Passador-Gurgel G, Hsieh W-P, Hunt P, Deighton N, Gibson G (2007) Quantitative trait transcripts for nicotine resistance in *Drosophila melanogaster*. *Nature genetics* 39(2): 264-268.
- Pili-Floury S, Leulier Fo, Takahashi K, Saigo K, Samain E *et al.* (2004) In vivo RNA interference analysis reveals an unexpected role for GGBP1 in the defense against Gram-positive bacterial infection in *Drosophila* adults. *The Journal of Biological Chemistry* 279(13): 12848-12853.
- Ramet M, Manfruelli P, Pearson A, Mathey-Prevot B, Ezekowitz RAB (2002) Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for E. coli. *Nature* 416(6881): 644-648.

- Rolff J (2002) Bateman's principle and immunity. *Proceedings Biological Sciences* 269(1493): 867-872.
- Sackton TB, Lazzaro BP, Clark AG (submitted) Genotype and gene expression associations with variation in the *Drosophila melanogaster* immune response. *PloS Biology*.
- Samakovlis C, Kimbrell DA, Kylsten P, Engstrom A, Hultmark D (1990) The immune response in *Drosophila*: pattern of cecropin expression and biological activity. *The EMBO Journal* 9(9): 2969-2976.
- Tanji T, Hu X, Weber ANR, Ip YT (2007) Toll and IMD pathways synergistically activate an innate immune response in *Drosophila melanogaster*. *Molecular and Cellular Biology* 27(12): 4578-4588.
- Tinsley MC, Blanford S, Jiggins FM (2006) Genetic variation in *Drosophila melanogaster* pathogen susceptibility. *Parasitology* 132(Pt 6): 767-773.
- Tzou P, Reichhart J-M, Lemaitre B (2002) Constitutive expression of a single antimicrobial peptide can restore wild-type resistance to infection in immunodeficient *Drosophila* mutants. *Proceedings of the National Academy of Sciences of the United States of America* 99(4): 2152-2157.
- Wang L, Weber ANR, Atilano ML, Filipe SR, Gay NJ *et al.* (2006) Sensing of Gram-positive bacteria in *Drosophila*: GGBP1 is needed to process and present peptidoglycan to PGRP-SA. *The EMBO Journal* 25(20): 5005-5014.
- Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA *et al.* (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genetics* 3(9): 1687-1701.
- Werner T, Borge-Renberg K, Mellroth P, Steiner H, Hultmark D (2003) Functional diversity of the *Drosophila* PGRP-LC gene cluster in the response to lipopolysaccharide and peptidoglycan. *The Journal of Biological Chemistry* 278(29): 26319-26322.
- Zuk M (1990) Reproductive strategies and disease susceptibility: an evolutionary viewpoint. *Parasitology Today* 6(7): 231-233.

CHAPTER 3

INFERRING QUANTITATIVE GENETIC PARAMETERS OF MARKERS ASSOCIATED WITH VARIATION IN IMMUNE FUNCTION IN *DROSOPHILA* *MELANOGASTER*

Introduction

It is now well established that many organisms harbor substantial natural variation for immune function (*e.g.* Ferrari *et al.* 2001; Niare *et al.* 2002; Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Tinsley *et al.* 2006; Sackton *et al.* submitted). Because pathogens are expected to exert strong selective pressure on immune system function, the maintenance of this variation is evolutionarily unexpected under a simple additive genetic model. A number of mechanisms have been proposed for the selective maintenance of variation for immune function, including costs of resistance (Antonovics and Thrall 1994; Kraaijeveld and Godfray 1997; Armitage *et al.* 2003; Yan and Severson 2003; Gwynn *et al.* 2005; McKean *et al.* 2008), antagonistic pleiotropy among resistance to diverse pathogens (Carius *et al.* 2001; Schmid-Hempel and Ebert 2003), and genotype-by-environment interactions (Mitchell *et al.* 2005; Lazzaro *et al.* 2008). Although selective explanations have received the most attention, polymorphism for immune system function could also represent the presence of weakly deleterious alleles maintained by mutation-selection balance. While considerable work has focused on testing these hypotheses, quantitative genetic characterizations of the underlying genetic basis for polymorphism in immune function have received comparatively little attention, with most studies characterizing variation in homozygous, inbred lines (Lazzaro *et al.* 2004; Lazzaro *et al.* 2008; Sackton *et al.* submitted) or focusing on individual components of the immune system

(Kraaijeveld and Godfray 1997; Cotter and Wilson 2002; Schwarzenbach *et al.* 2005; Luong and Polak 2007). However, the genetic architecture of resistance may greatly influence the tendency to maintain polymorphism, due for instance to overdominance or some forms of symmetric epistasis. Formal quantitative genetic analysis is a critical first step toward understanding the evolutionary dynamics of segregating genetic variation for immune system function, and thus shedding light on the role of selection in maintaining variation for immunocompetence.

Considerable previous work in *Drosophila* has helped to elucidate the molecular mechanisms of the immune response. Unlike vertebrates, insects do not have specific immunological memory, and thus lack antibody-mediated adaptive immunity to previously encountered pathogens. Instead, they use generalized, innate pathways that are deployed rapidly in response to bacterial and fungal infections (reviewed in Lemaitre and Hoffmann 2007). Foreign microorganisms are detected by conserved cell wall components such as PGN, which are recognized by circulating peptidoglycan-binding proteins (Steiner 2004; Aggrawal and Silverman 2007). This in turn leads to the activation of the *Toll* and *Imd* signal transduction pathways (named for major proteins in each pathway) and the translocation of NF- κ B transcription factors into the nucleus. These transcription factors upregulate a suite of several hundred downstream proteins (De Gregorio *et al.* 2001; Boutros *et al.* 2002; De Gregorio *et al.* 2002), including antimicrobial peptides, which function to directly kill foreign microorganisms (Imler and Bulet 2005).

Recent work in *Drosophila* has also begun to reveal candidate genetic markers that are significantly associated with the ability of the fly to resist internal growth of bacteria (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted), providing a unique opportunity to begin to assess the additive and dominance contributions to immune phenotypes of candidate functional polymorphisms. In this study, we focus on

a set of putative functional polymorphisms identified by a candidate gene association mapping approach in chromosome substitution lines of *D. melanogaster* (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted). These association mapping studies were carried out in homozygous chromosome substitution lines. While this approach allows rapid and repeatable measurement of immune phenotypes from many genetic lines that differ at known markers, homozygous chromosome substitution lines prevent any inference of the relative contribution of additive and dominance effects to phenotypic variation.

Here, we extend this previous work by constructing a series of seven diallel crosses to estimate the additive and dominance contributions of specific polymorphic markers (and the variation linked to them) to genetic variation in immune function after challenge with each of two different bacteria, *Serratia marcescens* and *Enterococcus faecalis*. For each diallel cross, we employed four lines carrying each of the two alleles at the marker of interest, and constructed the cross using those eight parental lines, in order to maximize our ability to make inferences about the selected marker. Using Bayesian regression, we estimate parameters of a genetic model for each targeted marker, and relate those estimates to predictions from previous studies and from models of the selective maintenance of variation in fitness traits. We provide evidence for both significant additive and dominance components of variation attributable to specific markers. Finally, we simulate data under standard genetic models to estimate our power to detect significant genetic effects, and show that while we have moderate power to detect relatively large additive effects, we may be underestimating the true extent of relatively small effects.

Materials and Methods

Drosophila stocks

The source material for the diallel design described below consists of two sets of previously described *D. melanogaster* chromosome substitution lines. In the first set, each line is homozygous for a second chromosome sampled from a natural population, with remaining chromosomes co-isogenic across all lines. These second chromosome substitution lines were originally collected from State College, PA, and were described initially in Lazzaro *et al.* (2004). We also examined a set of third chromosome substitution lines, also originally collected from State College, PA, and described in Fiumera *et al.* (2007). We have previously assayed immune response in the second chromosome lines (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006) and the third chromosome lines (Sackton *et al.* submitted). These lines have been genotyped for 127 second chromosome candidate immune polymorphism and 137 third chromosome candidate immune polymorphisms, respectively, as previously described (Lazzaro *et al.* 2004; Sackton *et al.* submitted).

Diallel design

In order to assay the additive and dominance effects of putative causal candidate regions selected based on our previous assays of immune function in these lines (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted), we designed targeted diallel crosses. Quantitative genetic crosses such as diallels have typically been used to partition the genetic variation in a particular trait into additive and non-additive components. In this instance, we have genotyped the parental lines at a number of candidate immune system genes, allowing us to estimate dominance and additive components of the regions linked to specific polymorphisms. We used the results of our previously published analyses of statistical associations between polymorphisms in candidate immune system genes and bacterial load to select four

Table 3.1. Marker information

Gene	Chr	SNP marker	MAF	Ancestral state	Effect Size (Ef)*	Effect Size (Sm)*	P-value (Ef)	P-value (Sm)	Marker Description**	Marker Selection
<i>GGBP1</i> ³	3	GGBP75D_3696	0.395	minor allele	-0.0179	-0.4925	0.943	0.001	A/G, in the only intron of <i>GGBP1</i> 3 bp	Highly significant in <i>Sm</i> and <i>Pr</i> infections [3]
<i>Imd</i> ^{1,2}	2	Marker 13475280	0.300	minor allele	-0.0143	0.3090	0.513	0.167	deletion/insertion 150 bp 5' of start	Highly significant in <i>Sm</i> infections [1] Significant in both <i>Sm</i> and <i>Ef</i>
<i>PGRP-LC</i> ³	3	PGRP_884	0.437	minor allele	0.5146	0.3414	0.050	0.034	C/T, 116 bp 5' of start	infections, large effect size [3]
<i>PGRP-SC1a</i> ^{1,2}	2	Marker 3771856	0.434	minor allele	-0.0050	-0.4240	0.266	0.002	G/A, 800 bp 3' of stop	Significant in <i>Sm</i> infections [1,2] Significant in <i>Sm</i> infections [1], L1
<i>SR-C11</i> ^{1,2}	2	Marker 7274899	0.138	major allele	0.3980	0.1370	0.146	0.678	27 bp deletion/insertion, intron 2	infections [2], large effect size
<i>Tehao</i> ^{1,2}	2	Marker 13423843	0.426	minor allele	-0.5510	0.2570	0.048	0.194	T/A, 778 bp 5' of start	Significant in <i>Ef</i> infections [2], large effect size
<i>Toll</i> ³	3	<i>Toll</i> _1186	0.080	major allele	0.9395	0.0496	0.098	0.889	T/C, in the 5' UTR	Large effect size [3]

1) Lazzaro *et al.* 2004, 2) Lazzaro *et al.* 2006, 3) Sackton *et al.* submitted

*Effect size is defined as the bacterial load of the genotype homozygous for the major allele minus the bacterial load of the genotype homozygous for the minor allele; thus, a positive value indicates the major allele has a higher bacterial load, and thus a less efficacious immune response. P-values are from prior tests of association in chromosome substitution lines. Estimates of effect size and P-values are derived from the data presented in Lazzaro *et al.* 2006 and Sackton *et al.* submitted.

**Minor allele is listed first

second chromosome polymorphisms and three third chromosome polymorphisms to target in this study. The selected SNPs that we targeted for a diallel cross are listed in Table 3.1, along with the effect size and *P*-value of the association between that SNP and immunocompetence from previous work (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted), and the rationale for selecting that SNP. While most markers were selected on the basis of highly significant associations identified by Lazzaro *et al.* (2006) or Sackton *et al.* (submitted), one marker in *Imd* was selected based on a highly significant association with *Serratia marcescens* load in Lazzaro *et al.* (2004) and one marker in *Sr-CII* was selected on the basis of a highly significant association with *Lactococcus lactis* as well as significant associations with other bacteria at linked markers in the same gene (Lazzaro *et al.* 2006).

For each selected polymorphism, we constructed a diallel cross by selecting four parental lines of each allele (for a total of eight parental lines), and crossing them in a complete half-diallel with no self crosses (Figure 3.1). This design increases our power to detect additive and dominance effects attributable to the targeted polymorphism, as we have in effect artificially constrained the allele frequency of the targeted polymorphism at 0.50, guaranteeing a balanced design. Since the parental lines are chromosome substitution lines, this design also minimizes background genetic effects, as only segregating variation on the chromosome on which the targeted marker resides will contribute to background genetic variation.

Diallel cross construction and phenotyping

To construct each diallel cross, we collected virgin females from the parental lines and established each cross in four replicate vials, with 5 males and 5 females (on average) per cross. All flies were reared on rich dextrose medium and kept at 25° on a 12h light / 12h dark cycle for the duration of the experiment. Approximately 3-5 days after emergence, we infected the offspring of each diallel cross with two different

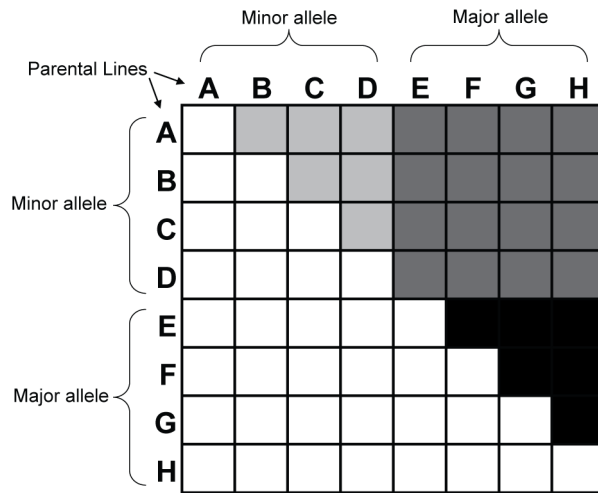


Figure 3.1.

Diallel crossing scheme used. Parental lines A-H were selected at random from the pool of second chromosome substitution lines (for crosses assaying markers in *PGRP-SC1a*, *Sr-CII*, *Tehao*, and *imd*) or the pool of third chromosome lines (for crosses assaying markers in *PGRP-LC*, *Toll*, and *GNBPI*), conditional on lines A-D being homozygous for the minor allele and lines E-H being homozygous for the major allele. The shaded boxes represent the F1 offspring assayed for bacterial load: light grey for minor allele homozygotes, grey for heterozygotes, and black for major allele homozygotes.

bacteria (*Serratia marcescens* and *Enterococcus faecalis*), with all infections for each cross/bacteria pair being done on a single day to minimize experimental error. Each cell of a given diallel cross was infected by three infectors in a complete-block design, with each infector infecting 12 flies of each sex from each line. Typically we obtained 3-4 replicates per line per infector per sex, for a total of approximately 20 replicates

per cell of the diallel. The bacterial strains used are as described in Lazzaro *et al.* (2006).

Bacterial load was measured as previously described (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted). Briefly, we pierced the thoraces of individual flies with a 0.1 mm dissecting pin (Fine Science Tools, Foster City, CA) coated in liquid culture (OD600 = 1.0±0.2) of the bacterium of interest, delivering an average of 4 x 10³ bacteria to each fly (not shown). To measure bacterial load, we homogenized same-sex trios of flies in 500 µl of sterile LB and then quantitatively plated the homogenates on standard LB agar plates using an Autoplate 4000 spiral plater (Spiral Biotech, Bethesda, MD). We incubated the plates overnight at 37°C and then estimated the concentration of viable bacteria in each homogenate using the colony counting system associated with the plater. Prior to plating, we diluted homogenates of *E. faecalis* 20-fold in sterile LB (10-fold for the *Sr-CII* cross), in order to correct for anticipated high bacterial loads. Our estimates of bacterial load per fly were log_e-transformed and normalized to correct for these dilutions before analysis.

Bayesian Regression

To fit genetic models and estimate dominance of each targeted SNP, we implemented a Bayesian regression in WinBUGS 1.4.3 (Lunn *et al.* 2000). As our primary goal is to estimate the additive and dominance components attributable to the targeted SNP, we focused on the following hierarchically centered model:

$$Y_{ijkl} \sim N(\mu_{ijk}, \sigma_e^2)$$

$$\mu_{ijk} \sim N(\mu_i + \beta^*j, \sigma_{inf}^2)$$

$$\mu_i \sim N(\text{geno}[i], \sigma_{gen}^2)$$

where Y_{ijkl} is the bacterial load phenotype for the l^{th} replicate infected by the k^{th} infector ($k=1-3$) of the j^{th} sex ($j=1,2$) from the i^{th} cross ($i=1-28$). The error variance is represented by σ^2_e , the variance attributable to infector is represented by σ^2_{inf} , and the background genetic variance (not attributable to the SNP effect) is represented by σ^2_{gen} . The effect of sex is represented by the β term, where female is coded as 1 and male as 2 (so β is equal to the average effect of being male). $\text{Geno}[i]$ represents the expected phenotype of the i^{th} cross based on genotype alone, which is defined by the following genetic model, where $\text{geno}[1]$ and $\text{geno}[3]$ are homozygotes, $\text{geno}[2]$ is the heterozygote and θ is the overall population mean:

$$\text{Geno}[1] = \theta - a$$

$$\text{Geno}[2] = \theta + d$$

$$\text{Geno}[3] = \theta + a$$

Under this genetic model, a represents the additive effect of the target SNP, and d represents the deviation from additivity in the heterozygote, with positive values indicating dominance of the $\text{geno}[3]$ homozygote and negative values indicating dominance of the $\text{geno}[1]$ homozygote. Values of d with a magnitude significantly larger than the magnitude of a suggest either overdominance or underdominance. In all cases, we define $\text{geno}[1]$ as the genotype composed of two copies of the minor (less frequent) allele, and $\text{geno}[3]$ as the genotype composed of two copies of the major allele; a thus represents the additive effect of replacing one copy of the minor allele with one copy of the major allele.

We assumed uninformative priors for all model parameters: normals with mean 0 and standard deviation 1000 for the fixed parameters θ , a , d , and β , and uniform distributions from 0 to 100 for the variance parameters σ^2_e , σ^2_{inf} , and σ^2_{gen} . All

parameters were estimated from 100,000 combined samples from two independent chains, after a burn in of 50,000 iterations and sampling every 10th iteration of the chain. All models converged very quickly (within 10,000 iterations).

Data simulation

We simulated data in R under two different assumptions about the structure of background genetic variation. In the first model, referred to as model 1, we do not distinguish between additive and non-additive background genetic effects, as in the model used for parameter estimation described above. In the second model, model 2, we assume a modified diallel model where the mean of the cross between the i^{th} sire and the j^{th} dam is described by:

$$\mu_{ij} \sim N(\text{geno}_{ij} + g_i + g_j, \sigma_{\text{sca}}^2)$$

where geno_{ij} is the genotype of the cross between the i^{th} sire and the j^{th} dam and follows the genetic model described above, g_i is the additive contribution of the i^{th} parent, and g_j is the additive contribution of the j^{th} parent, and σ_{sca}^2 is the variance attributable to non-additive effects of the cross (equivalent to the SCA variance in a traditional diallel model). Parent-of-origin effects are assumed to be absent. The additive effects (g) are assumed to be normally distributed with a mean of 0 and a variance equal to the GCA variance.

Results

Estimation of genetic contribution to variation

We selected seven putative functional polymorphisms for resistance to bacterial infection in *Drosophila*, based on previous candidate gene association studies (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted), and estimated dominance and additive effects attributable to each marker using a diallel design.

These seven markers include four in genes on the second chromosome (*Tehao*, *PGRP-SC1a*, *Sr-CII*, and *imd*), and three in genes on the third chromosome (*PGRP-LC*, *GNBPI*, and *Toll*). Full details of the markers chosen are presented in Table 3.1. For each marker, we selected four lines at random carrying each allele, giving a total of eight lines that were used as the parental stock for a diallel cross for each targeted marker (Figure 3.1; see Materials and Methods for details).

For each of these seven diallel crosses, we estimated the genetic contribution to variation in bacterial load after artificial infection using a Bayesian regression model. This model allows for a genetic model associated with the effect of the targeted polymorphism, as well as the estimation of background genetic effects that represent both the additive and non-additive contributions of other alleles to the bacterial load phenotype (see Materials and Methods for details). We fitted a standard genetic model in which the additive effect of the allele is estimated by the parameter a and the deviation of the heterozygote from the mean of the homozygous genotypes is estimated by the parameter d . We normalized the parameter a so that it represents the additive effect on phenotype of replacing one copy of the minor allele with one copy of the major allele. If a is positive, the major allele contributes to higher bacterial load, and thus each copy of the major allele an organism carries reduces immune function. If a is negative, the reverse is true, and each copy of the major allele the organism carries increases immune function.

In addition to the parameters of the genetic model, which are discussed in the next two sections, we estimated three variance parameters for each diallel cross (Table 3.2). One was a background genetic variance, representing both the additive and non-additive genetic contributions to variation in bacterial load that are attributable to genetic differences on either the second or third chromosome (depending on the cross)

Table 3.2. Estimated variance components

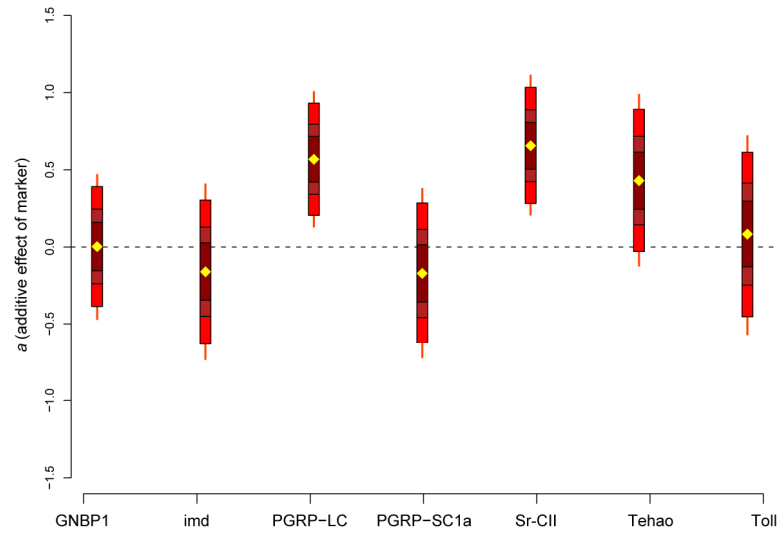
Bacterial challenge	Marker	Background Genetic Variance (Vgen)	Infector Variance	Residual Variance	Vgen / Vtotal (95% credible interval)
<i>E. faecalis</i>	<i>GNBP1</i>	0.2303	0.5299	6.70	0.0248 (0.0013, 0.0929)
<i>E. faecalis</i>	<i>imd</i>	0.3856	0.7082	7.56	0.0379 (0.0028, 0.1196)
<i>E. faecalis</i>	<i>PGRP-LC</i>	0.1631	0.3125	6.63	0.0172 (0.0007, 0.0759)
<i>E. faecalis</i>	<i>PGRP-SC1a</i>	0.2300	1.5530	7.76	0.0179 (0.0007, 0.0803)
<i>E. faecalis</i>	<i>Sr-CII</i>	0.1617	0.9169	5.99	0.0170 (0.0007, 0.0757)
<i>E. faecalis</i>	<i>Tehao</i>	0.2924	1.7490	6.95	0.0253 (0.0012, 0.1001)
<i>E. faecalis</i>	<i>Toll</i>	0.3962	0.3287	9.85	0.0309 (0.0020, 0.1070)
<i>S. marcescens</i>	<i>GNBP1</i>	0.1512	1.3550	3.43	0.0231 (0.0010, 0.0994)
<i>S. marcescens</i>	<i>imd</i>	0.1011	0.5165	2.19	0.0286 (0.0015, 0.1087)
<i>S. marcescens</i>	<i>PGRP-LC</i>	0.0711	0.3247	1.33	0.0341 (0.0021, 0.1174)
<i>S. marcescens</i>	<i>PGRP-SC1a</i>	0.0912	0.1601	1.06	0.0624 (0.0095, 0.1626)
<i>S. marcescens</i>	<i>Sr-CII</i>	0.1585	1.0460	2.53	0.0345 (0.0020, 0.1225)
<i>S. marcescens</i>	<i>Tehao</i>	0.1397	2.3670	3.52	0.0163 (0.0006, 0.0815)
<i>S. marcescens</i>	<i>Toll</i>	0.0734	0.1737	1.21	0.0437 (0.0044, 0.1294)

but not the genotyped marker being tested. The other two variance parameters are the variance attributable to the effect of different infectors; and the residual variance, a combination of the sampling variance (the error in estimating the bacterial load of a given pool of flies) and the environmental variance. While the estimated fraction of the total variance that is attributable to background genetic effects is relatively small (posterior mean estimates range from 0.022 to 0.068; Table 3.2), these estimates represent the proportion of variance explained by only about 40% of the genome. Furthermore, if considered an upper bound on the narrow-sense heritability conditional on sharing a genotype at the targeted marker, they are roughly consistent with previous estimates in *Drosophila melanogaster* (McKean *et al.* 2008).

Additive effects of genotyped markers on bacterial load

We estimated the posterior mean additive effect of each genotyped marker on bacterial load after artificial infection with either *E. faecalis* (Figure 3.2a) or *S. marcescens* (Figure 3.2b). For bacterial load measured after *E. faecalis* infection, two markers (in *PGRP-LC* and *Sr-CII*) have estimates of a where the 95% credible interval excludes 0, and a third marker (in *Tehao*) has an estimate of a for which the 70% credible interval excludes 0 (although the posterior mean value of a is opposite what we would predict based on the association testing results for this case). For bacterial load measured after *S. marcescens* infection, one marker (in *PGRP-LC*) has an estimate of a for which the 70% credible interval excludes 0, and one marker (in *PGRP-SC1a*) has an estimate of a for which the 50% credible interval excludes 0. For all other markers, the 50% credible interval of a contains 0, suggesting little evidence for a detectable additive effect at those markers, despite in some cases (*e.g.*, *GNBPI*) very strong evidence for a statistical association between the typed marker and bacterial load in previous studies (Table 3.1).

A. After *E. faecalis* challenge



B. After *S. marcescens* challenge

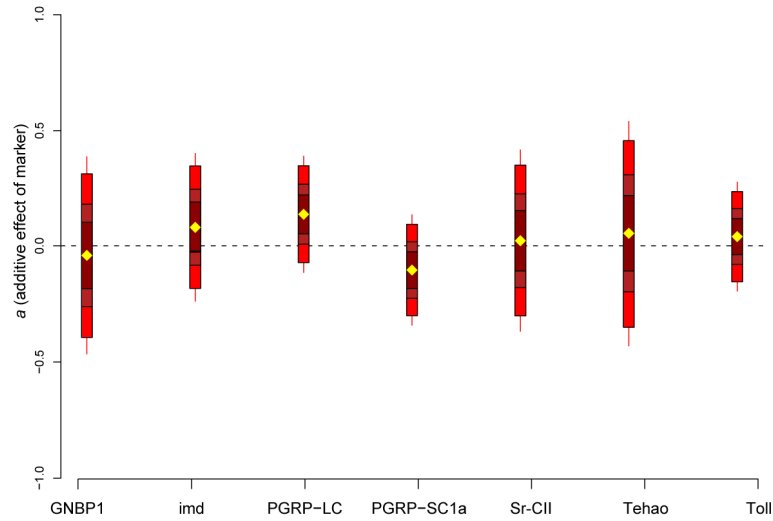


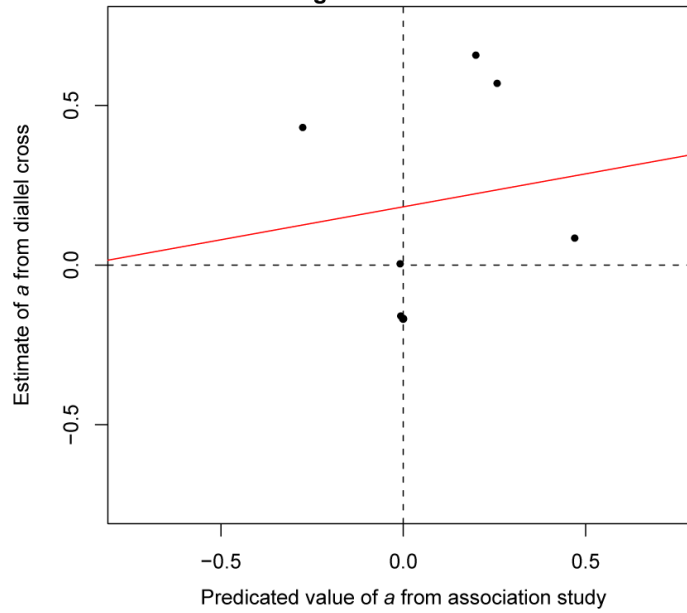
Figure 3.2.

Additive effects attributable to selected markers after challenge with A) *E. faecalis* or B) *S. marcescens*. Yellow diamonds represent posterior mean estimates, with the 50% (dark red), 70% (medium red), 90% (light red) and 95% (red line) credible intervals shown.

The estimated additive effect attributable to the marker at *Sr-CII* is of particular interest. The typed marker is a 27 bp deletion in the second intron of the gene, and while it is linked to several other SNP markers within the intron, LD decays rapidly outside the intron (Lazzaro *et al.* 2006), suggesting that the typed marker may be the causal allele or in tight linkage with the causal allele. None of the other markers in the second intron that were genotyped by Lazzaro *et al.* (2006) are perfectly associated with genotype (data not shown), suggesting that in order to estimate a significant effect, the causal site must be very tightly linked to the typed marker, or be the typed marker itself.

Based on the effect size of each marker in previous experiments, we can predict the expected value of a for each marker, assuming that the diallel experiments described here exactly replicate the previous results, as one-half the estimated effect size (Table 3.1). We find a strong and highly significant correlation between this expected value and the posterior mean of a estimated from our Bayesian regression for the crosses challenged with *S. marcescens* ($r^2 = 0.835$, $F_{1,5} = 25.31$, $P = 0.004$; Figure 3.3a), despite the relatively weak support for some associations detected in previous studies. However, after challenge with *E. faecalis*, we see very little correlation between the expected values of a based on previous studies and the posterior mean of a ($r^2 = 0.02$, $F_{1,5} = 0.10$, $P = 0.759$; Figure 3.3b). Furthermore, the root mean square error between the expected and posterior values of a are considerably smaller for the *S. marcescens* data than the *E. faecalis* data (*S. marcescens* median = 0.075, *E. faecalis* median = 0.311, Mann-Whitney U test $P = 0.053$). It is not clear why we might fail to recover the predicted additive effects after challenge with *E. faecalis*, especially given the apparent limited evidence for dominance components of genetic variation after challenge with this bacterium (see below), although it is worth noting that the error variance is substantially higher after infection with *E. faecalis* than after

A. After *E. faecalis* challenge



B. After *S. marcescens* challenge

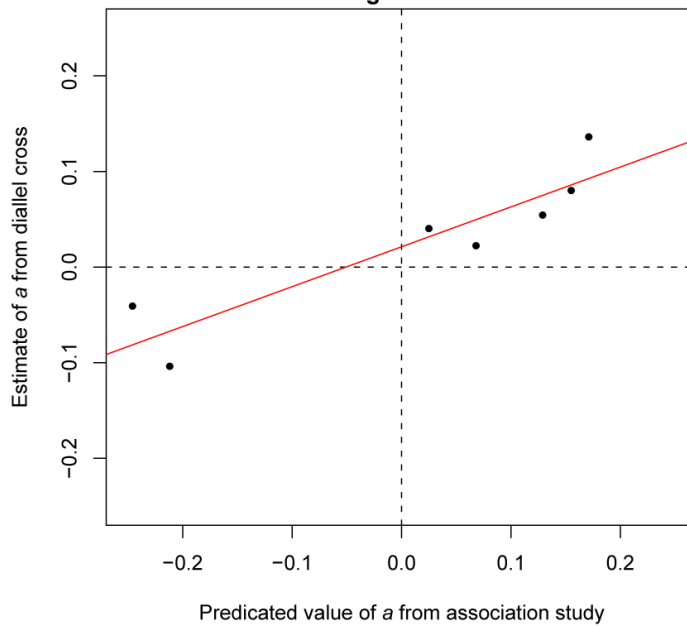


Figure 3.3.

Regression of predicated value of the additive effect of each marker against the observed estimate from the diallel crosses, on bacterial load after infection with A) *E. faecalis* ($r^2 = 0.02$) and B) *S. marcescens* ($r^2 = 0.835$).

infection with *S. marcescens* (median *E. f.* $\sigma^2_{\text{error}} = 6.95$, median *S. m.* $\sigma^2_{\text{error}} = 2.19$, Mann-Whitney U test $P = 0.0006$; Table 3.2).

Dominance component of selected markers

In addition to an additive effect, our genetic model includes a dominance effect as well, represented by the parameter d . In our parameterization, d represents the deviation of the heterozygote class from the expectation under additivity; a posterior estimate of d significantly different from zero provides evidence for a non-additive genetic component attributable to the targeted SNP. We find relatively little evidence for non-additivity in any of the diallel crosses after infection with *E. faecalis* (Figure 3.4a), although for two markers (in *Tehao* and *GNBPI*) the 50% credible interval of d excludes 0. After infection with *S. marcescens*, however, we find considerably more evidence for departures from additivity (Figure 3.4b): one marker for which 95% credible interval for d excludes 0 (in *imd*), one marker for which the 90% credible interval excludes 0 (in *Tehao*), one marker for which the 70% credible interval excludes 0 (in *GNBPI*), and two markers for which the 50% credible interval excludes 0 (in *PGRP-SC1a* and *Sr-CII*). Furthermore, for all five of these markers, the magnitude of the posterior estimate of d after challenge with *S. marcescens* is larger (although not always significantly so) than the magnitude of the posterior estimate of a , suggesting the possibility of extensive over- and under-dominance. In contrast, after challenge with *E. faecalis*, for five out of the seven markers the magnitude of the posterior estimate of d is smaller than the magnitude of the posterior estimate of a .

We considered the two cases with the most evidence for non-additivity (*imd* and *Tehao* in response to *S. marcescens* challenge) in more detail. There is no evidence for an additive effect on bacterial load attributable to the marker in *imd* (Figure 3.2b), and comparison of the posterior distribution of d to the posterior estimates of a suggests evidence for underdominance (Figure 3.4b). The posterior

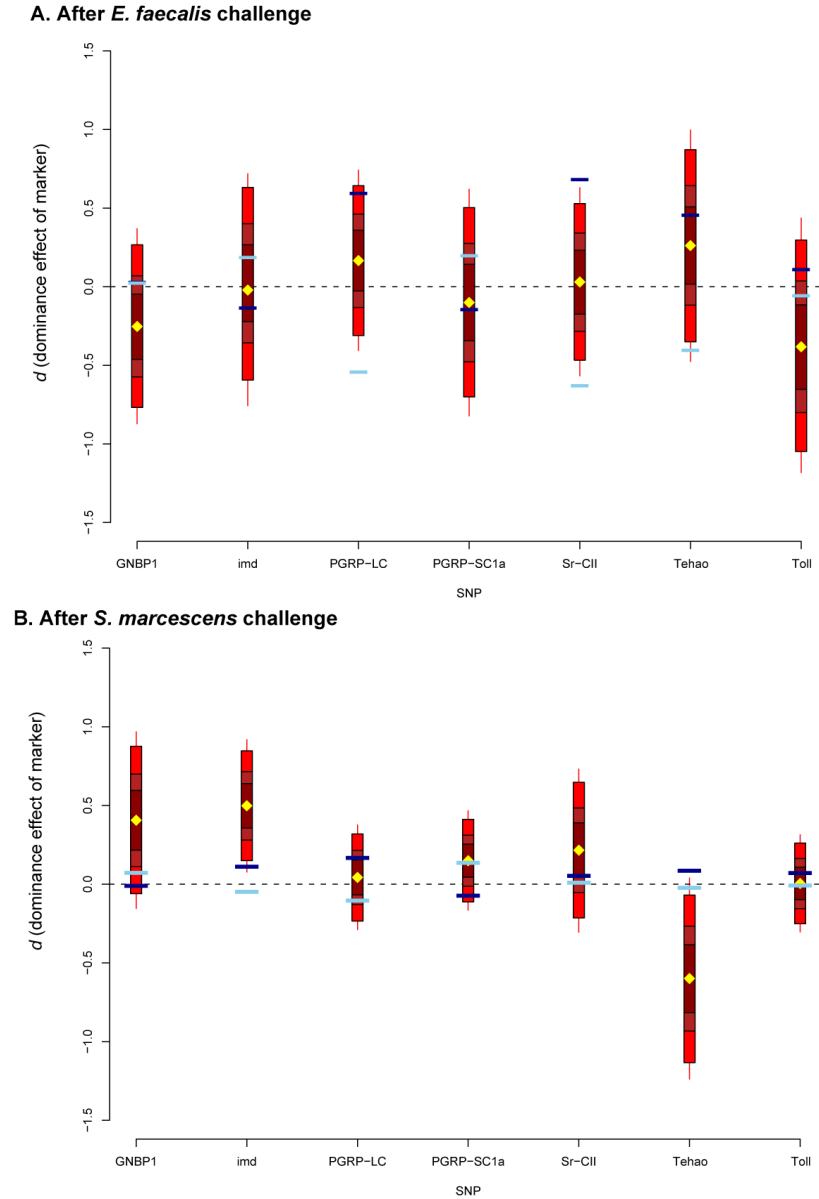


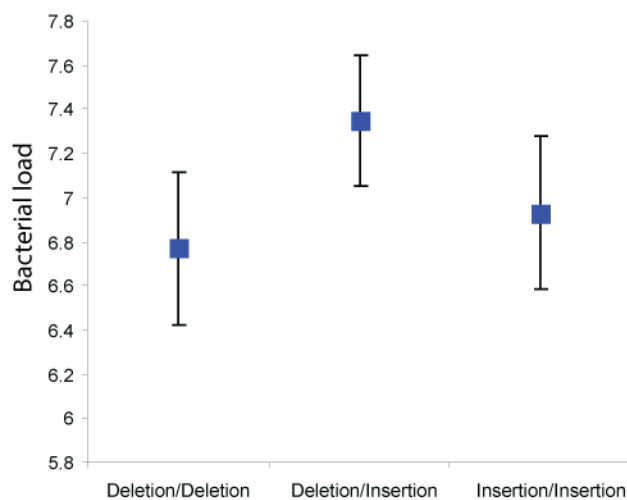
Figure 3.4.

Dominance effects (deviation of the heterozygote from the predicted phenotype under additivity) attributable to selected markers after challenge with A) *E. faecalis* or B) *S. marcescens*. Yellow diamonds represent posterior mean estimates, with the 50% (dark red), 70% (medium red), 90% (light red) and 95% (red line) credible intervals shown. Light blue bars indicate $-a$ and dark blue bar indicate $+a$.

mean estimates of the bacterial load after infection with *S. marcescens* of the three genotypic classes (Figure 3.5a) show the pattern clearly, with no apparent difference in bacterial load sustained after infection between the two homozygous classes, but a substantial increase in bacterial load sustained after infection in the heterozygote. It is important to note that these data do not necessarily imply that underdominance is being caused by the *imd* locus itself, as marginal underdominance can arise via epistatic interactions. Maintenance of underdominance as a stable polymorphism is typically difficult, although marginal underdominance generated by epistasis can exist as a stable polymorphism under some parameter conditions (Hastings 1981, 1982).

The marker in *Tehao* also appears to have no readily detectable additive effect on bacterial load phenotypes (Figure 3.2b), and comparison of the posterior distribution of d to the posterior estimates of a suggests evidence for overdominance (Figure 3.4b), which is supported by the posterior mean estimates of bacterial load of the three genotypic classes (Figure 3.5b). If overdominance is in fact acting at this locus, and if our measure of bacterial load is correlated with net reproductive fitness, it suggests a possible direct role for selection in the maintenance of polymorphism at *Tehao*. Notably, previous work has suggested that this locus may harbor antagonistically pleiotropic variation, in which the allele that is associated with decreased load after challenge with one bacteria is associated with increased load after challenge with a different bacteria (Lazzaro *et al.* 2006). While we do not find substantial support for antagonistic pleiotropy here, the repeated observation of complex interactions between genotype and pathogen at this locus implies a complex basis for phenotype and may increase the probability that selection can maintain stable polymorphisms at this locus.

A. *Imd* genotypes after infection with *S. marcescens*



B. *Tehao* genotypes after infection with *S. marcescens*

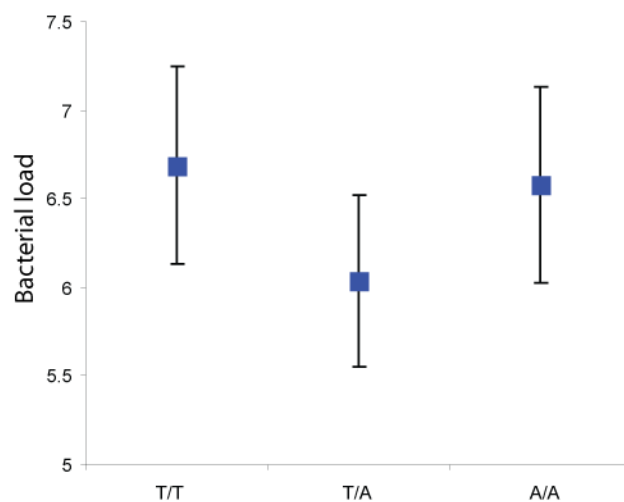


Figure 3.5.

A) Posterior means \pm 1 s.d. of each genotype of *Imd* after infection with *S. marcescens*. B) Posterior means \pm 1 s.d. of each genotype of *Tehao* after infection with *S. marcescens*.

Validation by simulation

In order to understand the statistical properties of the Bayesian regression framework we use to estimate parameters of the genetic model, we simulated data under two different models and with a series of parameter values for a and d . Both models assume the same genetic model, but differ in the assumptions about the structure of background genetic variation. In model 1, we do not partition background genetic variation into additive and non-additive components, whereas in model 2 we assume an additive background genetic variance distinct from the non-additive component (equivalent to a GCA and SCA component in a standard diallel). The full details of the simulations are described in the Materials and Methods.

Under both model 1 and model 2, we see no bias in our estimates of the parameters of the genetic model (data not shown). The mean of the posterior means across simulation replicates is very close to the true value of the parameter for both a and d . However, these simulations do suggest that our power to detect significant departures from zero for the parameters of our genetic model may be low. If we use the 95% credible interval of the estimated parameter values as a measure of significance, our false negative rate (fraction of cases where the 95% credible interval includes 0 when the true parameter value is non-zero) for the parameter combination $a=0.5$ and $d=0.0$ estimated from simulations under model 1 is 0.598; under model 2, it is 0.602. With smaller effect sizes, our false negative rate increases predictably (Table 3.3). While these results imply that we may miss effects of small size, the lack of bias in our parameter estimates even when simulating under a more complex model than the one we use to analyze the data suggests that the effects we do observe are probably robust.

Table 3.3. Simulations

model	True parameter values		N*	a		d	
	a	d		false negative rate**	error***	false negative rate**	error***
1	0	0	200	NA	0.035	NA	0.050
1	0.2	0	100	0.940	0.060	NA	0.030
1	0.5	0	1000	0.598	0.045	NA	0.036
1	0.2	0.2	100	0.920	0.050	0.920	0.020
1	0.5	0.5	100	0.720	0.060	0.820	0.040
2	0	0	200	NA	0.220	NA	0.120
2	0.2	0	100	0.840	0.210	NA	0.110
2	0.5	0	1000	0.602	0.193	NA	0.114
2	0.2	0.2	100	0.820	0.300	0.820	0.210
2	0.5	0.5	100	0.520	0.260	0.700	0.110

* number of simulations run

** fraction of tests where 95% credible interval includes 0 when the true value of the parameter is non-zero

*** fraction of tests where 95% credible interval excludes the true value of the parameter

Discussion

We have evaluated seven candidate markers, identified on the basis of candidate gene association studies for resistance to infection, using a diallel framework to estimate additive and dominance components of genetic variation in bacterial load sustained after challenge with two bacteria, *S. marcescens* and *E. faecalis*. By crossing parental lines with known genotypes and measuring the phenotype of their offspring, we are able to estimate the genetic effects attributable to specific candidate markers (and linked polymorphisms). Specifically, using a diallel crossing scheme allows us to extend the results of association studies conducted using homozygous chromosome substitution lines by facilitating the ability to disentangle additive and non-additive genetic contributions to bacterial load phenotypes.

Our results can serve as a test of the repeatability, in a more complex genetic context, of previously observed statistical associations between genotype and phenotype (Lazzaro *et al.* 2004; Lazzaro *et al.* 2006; Sackton *et al.* submitted). Here, our mixed results provide useful insights regarding the genetic architecture of immune function. On the one hand, we do replicate the predicted association between the marker in *PGRP-LC* and bacterial load after infection with *E. faecalis* and *S. marcescens*; this marker is the only one of the seven we selected with consistently large effects on bacterial load after infection with *S. marcescens* and *E. faecalis* in the original study (Sackton *et al.* submitted), suggesting that associations with large effects in multiple bacteria may be particularly robust to the differences between the experimental designs.

On the other hand, for several markers we estimate much smaller additive effects than expected, given the very strong evidence for statistical associations in the initial candidate gene screens. Furthermore, our estimates of a tend to be smaller than what we would predict based on previous work, especially after challenge with *S. marcescens*. While this could reflect the expected decline in effect size with repeated associations due to the winner's curse (Goring *et al.* 2001), it could also be a consequence of the heterozygous genomic context. Overall mean bacterial loads sustained tend to be higher in homozygous genetic lines, suggesting that these lines are generally worse at resisting infection, and may represent a sensitized assay for immune function. Finally, we have sampled many fewer lines in this assay, and depending on the strength of linkage disequilibrium between the causal site and the genotyped site, we may have sampled lines that carry the typed site but not the causal site.

Estimates of additive and dominance variation in complex traits can also illuminate hypotheses regarding the maintenance of variation in populations (Turelli

1984; Barton and Turelli 1989). In particular, the relative proportions of total genetic variance explained by additive (V_A) and dominance components (V_D) is predicted to vary under different models of the maintenance of genetic variation in populations (Charlesworth 1987; Charlesworth and Hughes 2000). Alleles maintained by pure balancing selection, by mechanisms such as heterozygote advantage, are not expected to contribute significantly to additive components of genetic variation (Charlesworth and Hughes 2000), and thus the ratio of V_D to V_A is predicted to be large for variation underlying traits maintained primarily by balancing selection. Variation maintained by recurrent deleterious mutation, on the other hand, is predicted to have a significant additive component, and thus a much smaller ratio of V_D to V_A is predicted under a mutation-selection balance model of the maintenance of variation. This contrast has been used (*e.g.*, Hughes 1997; Telonis-Scott *et al.* 2005) to attempt to disentangle the role of selection and mutation in maintaining polymorphism for fitness-related traits.

While the observation of substantial additive genetic variation can rule out the simplest forms of balancing selection, in which the heterozygote genotype has the highest fitness at equilibrium, it cannot rule out selective maintenance of genetic variation completely. Antagonistic pleiotropy, either between immune function and other fitness traits, or between immune function against different pathogens, can decouple overall organismal fitness from phenotypes at a particular component of fitness such as immunocompetence. Because this situation is predicted by a variety of trade-off models commonly proposed as mechanisms for the maintenance of variation in immune function, estimates of additive and dominance components of genetic variation for any given SNP cannot uniquely determine the role of selection in the maintenance of polymorphism. Furthermore, under some conditions genotype by environment interactions can result in the selective maintenance of largely additive genetic variation (Gillespie and Turelli 1989), as can certain models of selection in

fluctuating environments (*e.g.*, Burger 1999). Nonetheless, characterizing patterns of dominance and additive variation across loci is an critical step towards understanding the underlying genetic architecture of complex traits.

The patterns we observe across the seven loci we studied suggest that genetic variation for immune function at these loci cannot be simply categorized as either additive or non-additive. Assuming that there is a strong correlation between organismal fitness and immune function, we find some support for the hypothesis that selection may act directly to maintain variation via balancing selection in the observation that *Tehao* appears to harbor overdominant variation for bacterial load after challenge with *S. marcescens*. While the molecular mechanism for overdominance at *Tehao* is not clear, this result at least hints at the possibility of selective maintenance of genetic variation. We also find strong evidence for additive effects on bacterial load after challenge with *E. faecalis* attributable to markers in two loci, *PGRP-LC* and *Sr-CII*. Overall, we see substantially more dominance effects than additive effects of bacterial load after infection with *S. marcescens*, and the reverse pattern on bacterial load after infection with *E. faecalis*. The underlying cause of this pattern is unclear, but it does suggest complex differences in the genetic architecture of the immune response against different bacteria.

The most incisive prediction to distinguish selective maintenance of genetic variation from mutation-selection balance concerns the allele frequencies of the causal segregating variants: under most conditions, mutation-selection balance will lead to segregating low-frequency deleterious alleles, whereas trade-offs typically will lead to alleles segregating at intermediate frequencies (Barton and Keightley 2002). From our association studies we have estimates of the allele frequencies for every typed SNP. We can use this information, combined with estimates of dominance and additive genetic variation for particular alleles, to begin to predict what fraction of segregating

polymorphisms that affect immune phenotypes may be maintained by natural selection.

The most striking example of this is found at the marker in *Sr-CII*. As previously discussed, this marker is likely to be either the causal marker itself or very tightly linked to the causal marker, given the patterns of linkage disequilibrium in *Sr-CII* (data not shown). Our estimate of d is very small relative to our estimate of a for this marker, suggesting that most or all of the variation in bacterial load phenotype attributable to genotype at this marker is additive. However, contrary to the predictions of a mutation-selection balance model, the relatively low frequency derived allele is in this case the allele that is associated with greater resistance against *E. faecalis* infection (assuming that the ancestral genotyped allele is linked to the ancestral typed allele). This suggests that a simple mutation-selection balance model may not be sufficient to explain the extent of observed genetic variation in immunity, even for loci that harbor mostly additive variation.

Overall, our data underscore the complexity of the genetic architecture of resistance in *Drosophila*, which is not surprising given the complexity of host-pathogen interactions in time and space. We find evidence for both dominance and additive effects, as well as suggestive evidence that the relative importance of additive and dominance components of variation may differ among bacteria. While some markers appear to have strong, repeatable, mostly additive effects on bacterial load (*PGRP-LC*, *Sr-CII*), such effects do not appear to be the general trend, and non-additive genetic effects appear to be common. Further work will be needed to disentangle the contributions other factors, particularly epistasis and genotype by environment interactions.

REFERENCES

- Aggrawal K, Silverman N (2007) Peptidoglycan recognition in *Drosophila*. *Biochem Soc Trans* 35(Pt 6): 1496-1500.
- Antonovics J, Thrall PH (1994) Cost of resistance and the maintenance of genetic polymorphism in host-pathogen systems. *Proceedings of the Royal Society of London Series B* 257(1349): 105-110.
- Armitage SA, Thompson JJ, Rolff J, Siva-Jothy MT (2003) Examining costs of induced and constitutive immune investment in *Tenebrio molitor*. *J Evol Biol* 16(5): 1038-1044.
- Barton NH, Turelli M (1989) Evolutionary quantitative genetics: how little do we know? *Annu Rev Genet* 23: 337-370.
- Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. *Nat Rev Genet* 3(1): 11-21.
- Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Developmental Cell* 3(5): 711-722.
- Burger R (1999) Evolution of genetic variability and the advantage of sex and recombination in changing environments. *Genetics* 153(2): 1055-1069.
- Carius HJ, Little TJ, Ebert D (2001) Genetic variation in a host-parasite association: Potential for coevolution and frequency-dependent selection. *Evolution* 55(6): 1136-1145.
- Charlesworth B (1987) The heritability of fitness. In: Bradbury JW, Anderson MB, editors. *Sexual Selection: Testing the Alternatives*. John Wiley. pp. 21-40.
- Charlesworth B, Hughes KA (2000) The maintenance of genetic variation in life history traits. In: Singh RS, Krimbas CB, editors. *Evolutionary genetics from molecules to morphology*. Cambridge University Press. pp. 369-391.
- Cotter SC, Wilson K (2002) Heritability of immune function in the caterpillar *Spodoptera littoralis*. *Heredity* 88(4): 229-234.
- De Gregorio E, Spellman PT, Rubin GM, Lemaitre B (2001) Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 98(22): 12590-12595.

- De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. The EMBO Journal 21(11): 2568-2579.
- Ferrari J, Muller CB, Kraaijeveld AR, Godfray HC (2001) Clonal variation and covariation in aphid resistance to parasitoids and a pathogen. Evolution 55(9): 1805-1814.
- Fiumera AC, Dumont BL, Clark AG (2007) Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. Genetics 176(2): 1245-1260.
- Gillespie JH, Turelli M (1989) Genotype-Environment Interactions and the Maintenance of Polygenic Variation. Genetics 121(1): 129-138.
- Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. American Journal of Human Genetics 69(6): 1357-1369.
- Gwynn DM, Callaghan A, Gorham J, Walters KF, Fellowes MD (2005) Resistance is costly: trade-offs between immunity, fecundity and survival in the pea aphid. Proc Biol Sci 272(1574): 1803-1808.
- Hastings A (1981) Marginal underdominance at a stable equilibrium. Proceedings of the National Academy of Sciences of the United States of America 78(10): 6558-6559.
- Hastings A (1982) Unexpected Behavior in Two Locus Genetic Systems: An Analysis of Marginal Underdominance at a Stable Equilibrium. Genetics 102(1): 129-138.
- Hughes KA (1997) Quantitative genetics of sperm precedence in *Drosophila melanogaster*. Genetics 145(1): 139-151.
- Imler JL, Bulet P (2005) Antimicrobial peptides in *Drosophila*: structures, activities and gene regulation. Chemical Immunology and Allergy 86: 1-21.
- Kraaijeveld AR, Godfray HC (1997) Trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster*. Nature 389(6648): 278-280.
- Lazzaro BP, Scurman BK, Clark AG (2004) Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. Science 303(5665): 1873-1876.

- Lazzaro BP, Sackton TB, Clark AG (2006) Genetic variation in *Drosophila melanogaster* resistance to infection: a comparison across bacteria. *Genetics* 174(3): 1539-1554.
- Lazzaro BP, Flores HA, Lorigan JG, Yourth CP (2008) Genotype-by-environment interactions and adaptation to local temperature affect immunity and fecundity in *Drosophila melanogaster*. *PLoS Pathogens* 4(3): e1000025.
- Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annu Rev Immunol* 25: 697-743.
- Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10(4): 325-337.
- Luong LT, Polak M (2007) Costs of resistance in the *Drosophila-macroecheles* system: a negative genetic correlation between ectoparasite resistance and reproduction. *Evolution* 61(6): 1391-1402.
- McKean KA, Yourth CP, Lazzaro BP, Clark AG (2008) The evolutionary costs of immunological maintenance and deployment. *BMC Evol Biol* 8: 76.
- Mitchell SE, Rogers ES, Little TJ, Read AF (2005) Host-parasite and genotype-by-environment interactions: Temperature modifies potential for selection by a sterilizing pathogen. *Evolution* 59(1): 70-80.
- Niare O, Markianos K, Volz J, Oduol F, Toure A *et al.* (2002) Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. *Science* 298(5591): 213-216.
- Sackton TB, Lazzaro BP, Clark AG (submitted) Genotype and gene expression associations with variation in the *Drosophila melanogaster* immune response. *PloS Biology*.
- Schmid-Hempel P, Ebert D (2003) On the evolutionary ecology of specific immune defence. *Trends in Ecology & Evolution* 18(1): 27-32.
- Schwarzenbach GA, Hosken DJ, Ward PI (2005) Sex and immunity in the yellow dung fly *Scathophaga stercoraria*. *J Evol Biol* 18(2): 455-463.
- Steiner H (2004) Peptidoglycan recognition proteins: on and off switches for innate immunity. *Immunological Reviews* 198: 83-96.

- Telonis-Scott M, McIntyre LM, Wayne ML (2005) Genetic architecture of two fitness-related traits in *Drosophila melanogaster*: ovariole number and thorax length. *Genetica* 125(2-3): 211-222.
- Tinsley MC, Blanford S, Jiggins FM (2006) Genetic variation in *Drosophila melanogaster* pathogen susceptibility. *Parasitology* 132(Pt 6): 767-773.
- Turelli M (1984) Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theor Popul Biol* 25(2): 138-193.
- Yan G, Severson DW (2003) Dynamics of molecular markers linked to the resistance loci in a mosquito-Plasmodium system. *Genetics* 164(2): 511-519.

CHAPTER 4

DYNAMIC EVOLUTION OF THE INNATE IMMUNE SYSTEM IN *DROSOPHILA*

Introduction

Immune systems must constantly evolve in order to remain effective, due both to changes in the suite of pathogens to which they are exposed, and to the evolution of virulence mechanisms. These dynamics can result in a strong signature of adaptive evolution in genes involved in the immune response (Hughes and Nei 1988; Schlenke and Begun 2003). However, general patterns have been difficult to discern, as most studies have focused on a small number of genes in a few particular species. The recent complete genome sequencing of 12 *Drosophila* species (Clark *et al.* 2007), coupled with extensive molecular knowledge of the mechanisms of *Drosophila* immunity, provides an opportunity to dissect the evolutionary history of the annotated *D. melanogaster* immune system across the entire *Drosophila* genus.

Drosophila mount both cellular and cell-free, or humoral, immune responses to pathogens (Lemaitre and Hoffmann 2007). The cellular immune response consists of phagocytosis of microbes, and cellular encapsulation and melanization of larger parasites such as parasitoid wasp eggs, by differentiated populations of hemocytes (Meister and Lagueux 2003). The humoral immune response is initiated by the recognition of conserved microbe-specific molecules such as peptidoglycan, leading to the activation of signaling cascades and the nuclear translocation of the NF- κ B transcription factors Relish, dorsal, and DIF, which induce the transcription of antimicrobial peptides (AMPs) and other effectors (Hultmark 2003; Steiner 2004). While this response depends largely on the Toll and imd pathways (De Gregorio *et al.* 2002), other signaling cascades such as JAK/STAT and JNK appear to play

supplementary roles (Boutros *et al.* 2002; Agaisse and Perrimon 2004). Many of these diverse immune responses are analogous to the innate immune responses of mammals, employing many of the same components and regulatory pathways, although unlike mammals, insects such as *Drosophila* lack an antibody-mediated adaptive immune response (Silverman and Maniatis 2001; Evans *et al.* 2003).

Comparisons among the previously sequenced genomes of the dipterans *D. melanogaster*, *Anopheles gambiae*, and *Aedes aegypti*, and the hymenopteran *Apis mellifera* have revealed considerable variation in the size and diversity of immune-related gene families (Christophides *et al.* 2002; Evans *et al.* 2006; Waterhouse *et al.* 2007). Complete genome sequences are now available for 12 species in the genus *Drosophila*: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi* (Clark *et al.* 2007). The moderate divergence among these species (40 million years to the most recent common ancestor) provides considerable additional power for studies of molecular evolution, allowing tests for positive selection that are not possible with the dramatically more divergent genomes previously available. Furthermore, the sequenced *Drosophila* species span a wide range of diverse habitats and ecologies, including tropical rain forest species (*D. erecta*, *D. yakuba*), island endemics (*D. sechellia*, *D. grimshawi*), cosmopolitan human commensals (*D. melanogaster*, *D. simulans*), and cactophilic desert species (*D. mojavensis*) (Markow and O'Grady 2007). *Drosophila* breed and lay eggs in rotting plant and fungal material, exposing them to a wide range of pathogens in these septic environments, including viruses, bacteria, fungi, protozoans, nematodes, and parasitic wasps.

In this study, we annotate orthologs and paralogs of characterized and candidate immune system genes across the genus *Drosophila*. We analyze patterns of

gene family expansion and contraction in all 12 sequenced species and identify the origination of evolutionary novel immune system genes. Using likelihood-based models of molecular evolution, we test for positive selection across immune-related genes in the *melanogaster* group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*), and identify the protein domains that are the most likely targets of adaptive evolution.

Results

Annotation of immunity proteins in the genus Drosophila

We used an initial set of 245 *D. melanogaster* immune-related proteins to identify and manually curate 2501 candidate homologs in the remaining *Drosophila* species (described in detail in Sackton *et al.* 2007). For many of our analyses, we grouped genes based on molecular functions: “recognition genes” that encode pathogen surveillance proteins (*e.g.*, peptidoglycan recognition proteins (PGRPs) and phagocytic receptors such as eater); “signaling genes” that encode proteins in immune-related signaling pathways (*e.g.*, Toll and imd); and “effector genes” that encode proteins that directly inhibit pathogen growth and survival (*e.g.*, AMPs). Any broad functional classification is necessarily subjective to some degree, and some proteins could plausibly be assigned to multiple categories (for instance, some recognition proteins also initiate signal transduction). Furthermore, the molecular functions of many candidate immune genes are currently inferred only by sequence similarity, resulting in multiple equally plausible classifications of *D. melanogaster* immune-related proteins. Therefore, we have conducted our analyses using several classification schemes that are modified from the one presented here, either by including only the subset of genes with high-confidence functional annotations, or by using alternative functional categories. These modified classification schemes do not substantially change our conclusions (data not shown).

Patterns of gene conservation across 12 *Drosophila* species

To initially assess gene conservation across the 12 sequenced species of *Drosophila*, we assigned homology patterns to one of three classes: “single-copy orthology” for genes conserved as single-copy orthologs in all 12 species; “conserved paralogy” for genes that vary in copy number across the phylogeny but that are

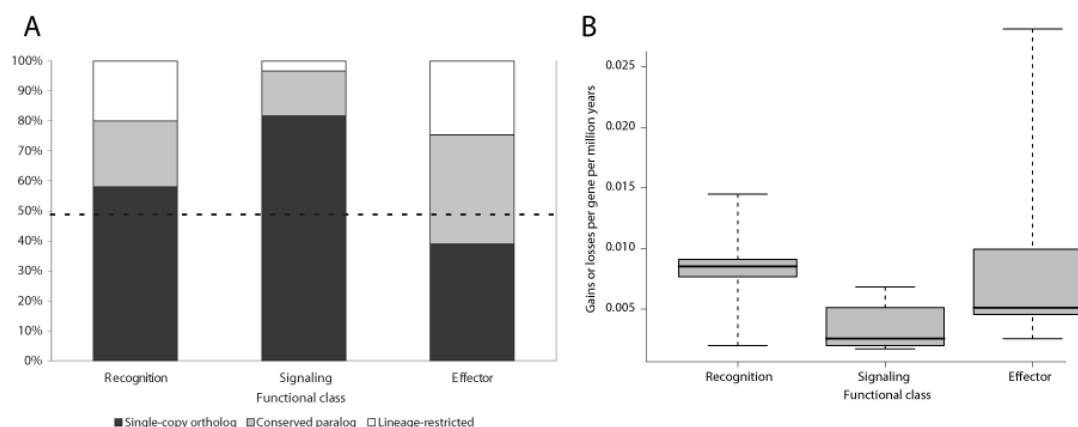


Figure 4.1

Variation in patterns of homology among immune-system genes. A) Proportion of each functional class assigned to each homology class. The dashed line is the fraction of the entire genome estimated to be in the single-copy ortholog class. B) Box plot of the estimated rate of gene turnover among genes in gene families for each functional class.

inferred to have been present in the common ancestor of Drosophilids; and “lineage-restricted” for genes that have arisen since that common ancestor. The proportion of genes in each homology class varies significantly among recognition, signaling, and effector classes, with the highest fraction of single-copy orthologs in the signaling class and the lowest in the effector class ($\chi^2 = 41.13$, d.f. = 4, $P = 2.53 \times 10^{-8}$; Figure 4.1A). Furthermore, only the effector class has a deficit of single-copy orthologs

relative to the genomic average (Figure 4.1A). This is not an artifact of the general pattern that short proteins are less likely to be single-copy orthologs, as AMPs (N=20) have significantly fewer single-copy orthologs than a control set of peptides (N=2878) of similar length ($\chi^2 = 14.92$, d.f. = 2, $P = 5.76 \times 10^{-3}$).

The variation in proportions of genes in each homology class implies variation in the rate of gene duplication and loss among functional classes of immune system genes. We used a recently developed maximum-likelihood model of birth-death evolution in gene families (Hahn *et al.* 2005) to estimate λ , the rate of gene turnover (duplications and losses) per million years. The distribution of λ varies among functional classes (Kruskal-Wallis test; $P = 0.012$), even when only gene families with at least one duplication or loss ($\lambda > 0$) are considered (Figure 4.1B; Kruskal-Wallis test; $P = 0.038$).

A prototypical example of the rapid changes in copy number among effector proteins is the cecropin gene family, a family of cationic peptides with antimicrobial activity against Gram⁺ bacteria, Gram⁻ bacteria and fungi. Cecropin homologs have been identified in all major endopterygote insect orders except *Hymenoptera*, and in many cases appear to be organized in a single genomic cluster (Hultmark 1993). As expected, we find a syntenically conserved cecropin cluster in all 12 *Drosophila* species in this study. There appear to have been at least four independent expansions of this cluster within the *Sophophora* subgenus, three within the *Drosophila* subgenus, and at least two independent losses within the *melanogaster* group. In principle, paralogous gene conversion can create a phylogenetic pattern similar to that expected from gene duplication and deletion. However, previous studies have found no evidence of gene conversion among cecropin genes in *D. melanogaster* (Clark and Wang 1997; Ramos-Onsins and Aguade 1998), and changes in gene order and orientation among species suggest rapid turnover, not gene conversion. This pattern of

rapid gene turnover with many independent expansions is common, if less extreme, in other effector and recognition gene families, in sharp contrast to signaling genes, where rates of gene duplication are dramatically lower.

Evolutionary novelties in the Drosophila immune system

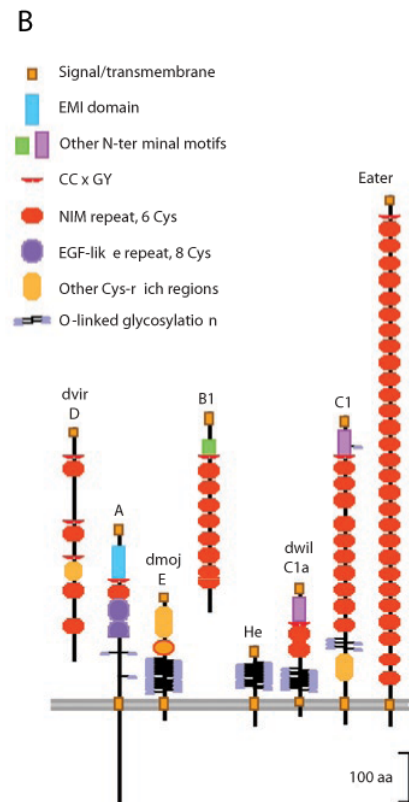
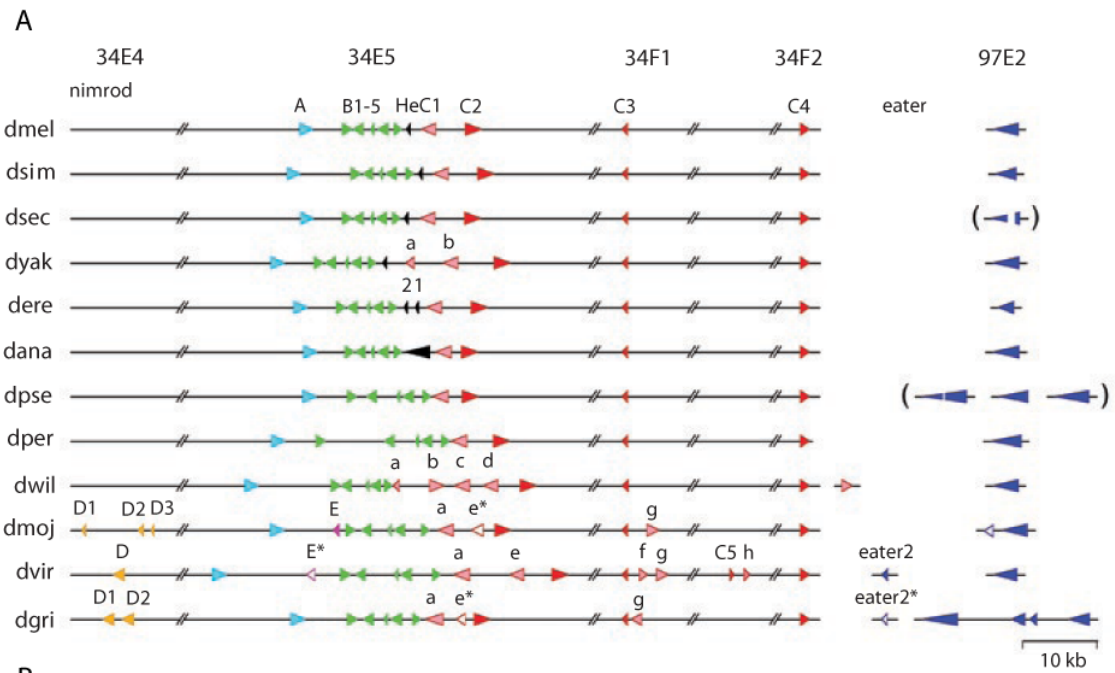
Comparisons among mosquitoes, fruit flies, and honeybees have identified lineage-specific genes encoding both recognition and effector proteins, suggesting the emergence of evolutionary novelties in the insect immune system (Christophides *et al.* 2002; Evans *et al.* 2006; Waterhouse *et al.* 2007). Based on the phylogenetic pattern of gene presence and absence in lineage-restricted gene families within *Drosophila*, we find evidence for the emergence of evolutionary novelties among recognition and effector gene families over roughly an order of magnitude shorter time scales. In contrast, the complement of signaling proteins in the immune system appears to be quite stable over the 40 million years of evolution since the root of the *Drosophila* genus, consistent with observations from more distant comparisons within insects (Christophides *et al.* 2002; Evans *et al.* 2006; Waterhouse *et al.* 2007).

Although the complement of proteins (such as PGRPs) that recognize microbe-specific molecules is essentially constant throughout the genus *Drosophila*, this is not the case for gene families thought to encode phagocytosis receptors. Of particular interest is the family that includes the eater and nimrod proteins, putative phagocytosis receptors characterized by a unique type of EGF-like repeat, the NIM repeat (Kocks *et al.* 2005; Kurucz *et al.* 2007). Members of this family, particularly *eater* and *nimrod C1* (*nimC1*), have independently expanded in several species (Figure 4.2A).

The *Hemese* (*He*) gene is located within the nimrod cluster and is also expressed in the hemocyte plasma membrane, but lacks NIM repeats and instead has a short serine and threonine-rich O-glycosylated extracellular domain (Kurucz *et al.* 2003). *He* homologs are not detectable outside the *melanogaster* group (Figure 4.2A),

Figure 4.2

Schematic map of the *nimrod* and *eater* genes in the 12 *Drosophila* species. A) Genomic organization of the *nimrod* genes and *eater*. Different classes of genes are color coded: *nimA*, light blue; *nimB*, green; *nimC*, red; *nimD*, orange; *nimE*, violet; *He*, black; *eater*, dark blue. Pseudogenes are denoted by empty symbols and are marked with an asterisk. *He* paralogs are numbered, and the *nimC1* paralogs are marked with lower-case letters. The sequence assembly is incomplete for the *eater* genes in *D. sechellia* and *D. pseudoobscura*. *dmel*, *D. melanogaster*; *dsim*, *D. simulans*; *dsec*, *D. sechellia*; *dyak*, *D. yakuba*; *dere*, *D. erecta*; *dana*, *D. ananassae*; *dpse*, *D. pseudoobscura*; *dper*, *D. persimilis*; *dwil*, *D. willistoni*; *dmoj*, *D. mojavensis*; *dvir*, *D. virilis*; *dgri*, *D. grimshawi*. B) Domain organization in the major classes of *nimrod* and *eater* proteins. Except for NimC1a, NimD and NimE, the examples shown are from *D. melanogaster*.



although one *nimCI* paralog in *D. willistoni*, *nimCIa*, has a similarly serine/threonine-rich region and a reduced number of NIM repeats (Figure 4.2B). A likely model is that the *He* gene originated from a truncated *nimCI* paralog that has lost all NIM repeats; *nimCIa* in *D. willistoni* may therefore represent a potential *He* analog. Within the nimrod family, *He* is not the only apparent evolutionary novelty: the Nimrod D subfamily appears restricted to the *Sophophora* subgenus, and the Nimrod E subfamily to the *D. virilis/D. mojavensis* clade (Figure 4.2A). Interestingly, the class C scavenger receptors (SR-Cs) in the *melanogaster* subgroup (a family of proteins related to SR-CI, a scavenger receptor expressed in the hemocyte plasma membrane and implicated in the phagocytosis of bacteria; Ramet *et al.* 2001) also appear to have diversified by partial or truncated duplications (Sackton *et al.* 2007). Novel genes often arise from rearrangements, truncations, and fusions of existing genes (Long *et al.* 2003), and the fixation of these novel genes may be a common mechanism to generate diversity in recognition proteins.

Apparent evolutionary novelties also exist in the effector class. The most striking example is the seven-member drosomycin antifungal peptide family, although we see a similar pattern for less-well-characterized effector protein families, such as the Turandots (Sackton *et al.* 2007). Homologs of drosomycin have previously been identified within the *melanogaster* and *ananassae* subgroups (Jiggins and Kim 2005), and in *D. triauraria*, a member of the closely-related *montium* subgroup (Daibo *et al.* 2001). The genomic arrangement of the drosomycin family is conserved and each drosomycin ortholog is monophyletic in the *melanogaster* subgroup; several rearrangements disrupt the drosomycin cluster in *D. ananassae*, suggesting independent expansion within the *ananassae* and *melanogaster* subgroups. Despite strong conservation of amino acid sequence among the *D. melanogaster* drosomycins, BLAST searches against the newly sequenced *Drosophila* genomes (and other

previously sequenced insect genomes) failed to identify putative homologs more distant than *D. ananassae*.

Surprisingly, we found drosomycin-like sequences in EST databases from three different coleopteran species (*e.g.* CB377292, DV767586, CV160723). Given the lack of drosomycins in any completely sequenced non-*Drosophila* insect genome, it is possible that these beetle ESTs represent contaminants or microbial products. However, if these are actually genuine beetle drosomycins, we suggest at least three possible explanations: drosomycins have been independently introduced in the *Drosophila* and/or coleopteran lineages by horizontal gene transfer (perhaps by *Wolbachia*; Hotopp *et al.* 2007); drosomycins have been lost independently in most flies and several other insect orders; or drosomycins have arisen at least twice by convergent evolution, perhaps from a defensin-like precursor.

These instances of apparent lineage-specific gains of known and putative immune-related proteins in *Drosophila*, combined with the known diversity of AMPs across insects (Bulet *et al.* 1999; Evans *et al.* 2006), suggest that there remain novel immune components to discover in *Drosophila* species outside the *melanogaster* group. While the core immune signaling pathways are deeply conserved as single-copy orthologs, there appears to be considerable flexibility in the inputs and outputs of the system that allows novel components to be integrated into the immune response over evolutionarily short time scales.

Patterns of positive selection in innate immune genes

We used codon substitution models of molecular evolution, implemented in the software package PAML (Yang 1997), to estimate ω (d_N/d_S , the relative rate of nonsynonymous to synonymous substitution) and infer patterns of positive selection (see Sackton *et al.* 2007 for further details). These models require accurate nucleotide alignments and become less reliable at high synonymous divergence, limiting our

analysis to the six species in the *melanogaster* group. To test for positive selection we compared the likelihood of the data under a model that requires a subset of codons to have $\omega > 1$ (a pattern predicted only when adaptive fixations have occurred) to the likelihood of the data under a model that does not allow such codons (Yang *et al.* 2000). Any gene that rejects the null model has some number of codons that have experienced significantly more nonsynonymous substitutions across the tree than expected. We used a false discovery rate (FDR) of 5%, unless otherwise noted, to correct for multiple testing (Storey and Tibshirani 2003).

Immune system genes evolve more rapidly than other genes

We compared ω estimated under the simplest model (a single ω per gene) between immune genes in this study and all single-copy orthologs in the *Drosophila* protein-coding genome (Clark *et al.* 2007, Larracuente *et al.* 2008). Immune genes are significantly less conserved than the set of all single-copy orthologs in the *melanogaster* group (immune genes: median $\omega = 0.080$, N=226; all single-copy orthologs: median $\omega = 0.064$, N=8510; $P = 1.43 \times 10^{-5}$, Mann-Whitney U). This pattern does not appear to be the result of biases introduced by the manual curation of immunity genes compared to the computational curation of the whole-genome dataset, as the results are qualitatively identical when only computationally curated immunity gene models are included. This elevated ω in immune genes is likely driven by adaptive evolution, as 514 of 8510 single-copy orthologs in the *melanogaster* group (6.0%) show evidence for positive selection after multiple test correction (Clark *et al.* 2007, Larracuente *et al.* 2008), compared to 23 of 226 immunity genes (10.2%), a difference that is significant by Fisher's Exact Test (FET) ($P = 0.016$; Figure 4.3). The strength of this effect depends slightly on what genes are classified as “immunity” for this analysis (data not shown).

Within immune system genes, the proportion of positively selected genes

differs among recognition, effector, and signaling classes (Figure 4.3). Compared to

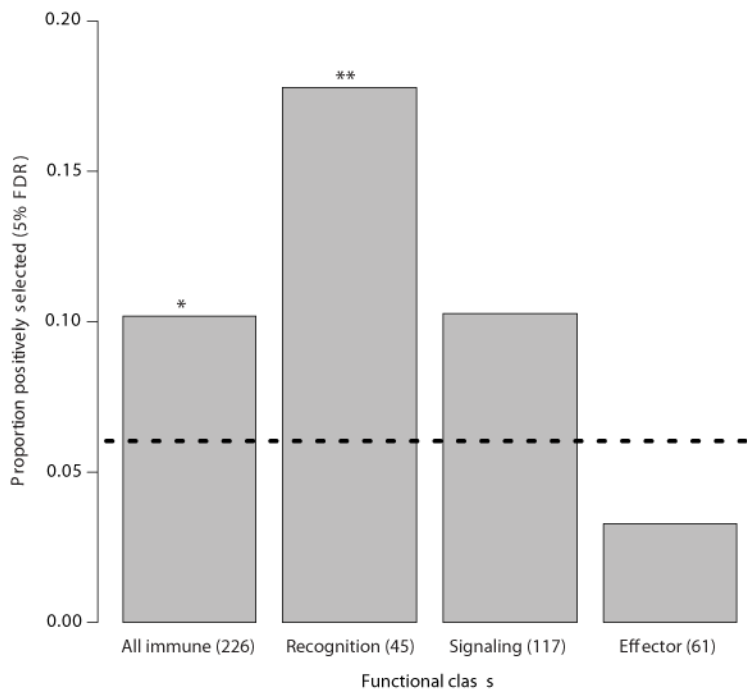


Figure 4.3

Variation in positive selection among immune-system genes. Proportion of positively selected genes (at a 5% false-discovery rate) within the immune system as a whole and in each functional class. Numbers in parentheses indicate the number of genes in each class. The dotted line is the estimate of the fraction of positively selected genes (at a 5% false discovery rate) among all single-copy orthologs, from ref. 3. Asterisks indicate a significant difference from the genomic fraction (FET: *, $0.01 < P < 0.05$; **, $P < 0.01$). Among functional classes, the proportion of positively selected genes varies significantly ($P = 0.046$, χ^2 -test).

the genomic single-copy ortholog dataset, genes that encode recognition proteins are significantly more likely to show evidence of positive selection (17.8% vs. 6.0%, $P = 0.005$, FET), a result that is robust to a number of different alternative classifications

of immune-system genes. Genes that encode signaling proteins trend towards excess positive selection relative to the genomic set (10.3% vs. 6.0%, $P = 0.076$, FET), and genes that encode effector proteins are less likely, although not significantly so, to be in the positive selection class (3.3% vs. 6.0%, $P=0.585$, FET). Although signaling proteins are more likely to have pleiotropic non-immune functions than recognition or effector proteins, differences in the degree of immune-specificity among functional classes do not appear to account for the variation in positive selection that we observe.

Positive selection drives the evolution of recognition proteins

Strikingly, of the 10 recognition genes ascertained to be evolving under positive selection (with a 10% FDR), two have been directly shown to participate in phagocytosis of foreign microorganisms (*NimC1*, Kurucz *et al.* 2007; *TepII*, Stroschein-Stevenson *et al.* 2006), and the seven of the remaining eight are homologous to proteins involved in phagocytosis in *Drosophila* or mammals (TEPs: *TepI*, *TepIV*; Nimrods: *NimB1*, *NimB4*; CD36 homologs: *crq*, *CG31217*, *emp*). Furthermore, two additional experimentally identified phagocytic receptors (*eater*, Kocks *et al.* 2005; *peste*, Philips *et al.* 2005) have some evidence for positive selection (*eater* nominal $P = 0.019$; *peste* nominal $P = 0.019$). In contrast, only *PGRP-LC* and *PGRP-LB* among the genes encoding PGRPs or Gram-negative binding proteins (GNBPs) have any evidence for positive selection. This excess of positive selection in putative phagocytosis genes is significant ($P = 0.034$; FET). One possible hypothesis for this difference is that the molecules recognized by PGRPs and GNBPs (peptidoglycan and β -glucan) are evolutionarily static, and thus unlikely to trigger coevolutionary arms races. In contrast, the targets of phagocytosis receptors may be more variable in structure, and thus more likely to lead to bouts of host-pathogen coevolution.

Of the fourteen recognition genes with evidence for positive selection at a nominal α of 0.05, we have reasonable hypotheses for the domain that might interact directly with pathogens for six of them (*TepI*, *TepII*, *TepIV*, *NimC1*, *eater*, and *PGRP-LC*). The TEP proteins are members of the $\alpha 2$ -macroglobin superfamily, and contain a hypervariable region that is likely important for interactions with pathogens (Blandin and Levashina 2004). Nimrod C1 and eater are both type I membrane proteins characterized by a large number of NIM repeats and a more divergent N-terminal region. The 200 N-terminal amino acids have been experimentally determined to be sufficient for bacterial binding in eater (Kocks *et al.* 2005), and likely have a similar function in NimC1. Both molecular and structural data show that the PGRP domain is required for binding to peptidoglycan in PGRP-LC (Chang *et al.* 2005; Chang *et al.* 2006). Using Bayesian estimates of the probability of positive selection for each codon in these six proteins (Yang *et al.* 2005), we find that codons in these “pathogen interaction domains” are significantly more likely to evolve by positive selection than codons outside these domains (Table 4.1), suggesting that adaptive evolution of these *Drosophila* recognition proteins is driven by interactions with pathogen-associated molecules.

One of these recognition proteins, PGRP-LC, is alternatively spliced in *D. melanogaster* to produce three isoforms with different PGRP domains attached to the same cytoplasmic domain (Werner *et al.* 2003). All three splice forms (PGRP-LCa, -LCx, and -LCy), are conserved in all 12 species, although we only find evidence for positive selection in the PGRP-LCa isoform, particularly in the PGRP domain (Table 4.1; Figure 4.4A). Two of the putative positively selected sites in PGRP-LCa (Ile-444 and Asn-445) are part of an insertion (relative to PGRP-LCx and PGRP-LCy) that induces a structural change leading to an altered binding properties (Chang *et al.* 2005; Chang *et al.* 2006). This two amino acid insertion is present in the five species of the

Table 4.1. Distribution of positively selected sites among recognition proteins.

Gene	Domain	Positively Selected Sites (1)	Total Sites	P-value (2)
PGRP-LCa	PGRP	4	166	
	Rest	1	354	0.0379
Eater	N-terminal	2	199	
	Rest	2	650	0.2354
NimC1	N-terminal	7	199	
	Rest	2	429	0.0057
Receptors Pooled	"Pathogen interaction domain"	13	564	
	Rest	5	1433	0.0001
TepI	Hypervariable	8	63	
	Rest	78	1298	0.0552
TepII	Hypervariable (3)	6	200	
	Rest	21	1350	0.1473
TepIV	Hypervariable	4	47	
	Rest	9	1445	0.0005
Teps Pooled	Hypervariable	18	310	
	Rest	108	4093	0.0038
All Pooled	"Pathogen interaction domain"	31	874	
	Rest	113	5526	0.009328

(1) Any site with a Bayesian posterior probability of positive selection greater than 0.75 is considered a "positively selected site"

(2) Calculated by Fisher's Exact Test

(3) Includes all splice for

melanogaster subgroup, but not in any more distant species (Figure 4.4B). It thus appears that the novel structural conformation induced by this insertion is evolutionary recent, and that selection may have acted to fine-tune the modified structure for improved stability, binding affinity, or some similar property.

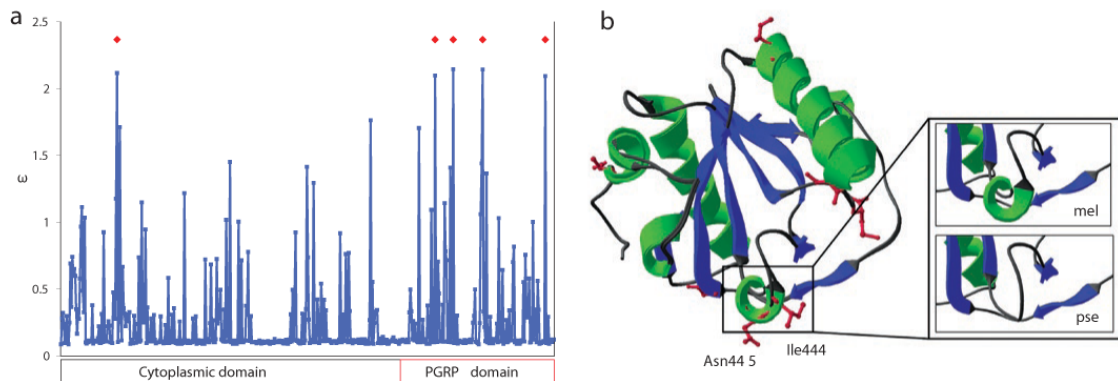


Figure 4.4

Positive selection in PGRP-LCa. (a) Variation in ω among codons of PGRP-LCa. The domain structure of the protein is represented on the x axis. Red diamonds mark codons with a Bayesian posterior probability of positive selection >0.75 . No such positively selected sites are found in the alternative exons encoding PGRP-LCx and –LCy domains. (b) Structural model of the PGRP-LCa domain from *D. melanogaster* (mel) and *D. pseudoobscura* (pse). The *D. pseudoobscura* sequence is threaded onto the *D. melanogaster* structure using Swiss-PDB Viewer with the default settings. Both structures are color-coded by secondary structure. Side chains of positively selected sites (posterior probability >0.50) are shown in magenta, with the structure-altering residues Asn444 and Ile445 labeled. The insert shows the primary divergent region between the two structures.

Signal modulation proteins evolve more rapidly than signal transduction proteins

Although signaling genes overall show only a non-significant trend towards excess positive selection relative to genomic averages, a different pattern emerges

when the signaling class is divided into genes encoding proteins with a modulation function and genes encoding proteins with a signal transduction function. Six out of 26 modulation proteins show evidence for positive selection (23.8%), significantly more than either the genomic average (6.0%; FET $P = 0.004$) or signal transduction proteins (6.4%; FET $P = 0.0217$). Modulation proteins also have a very different pattern of copy-number conservation: 51.7% are found as single-copy orthologs, compared to 88.4% of signal transduction proteins. These differences may result from modulation proteins occupying less central, and therefore less constrained, positions in innate immune signaling networks.

Positive selection in the Relish cleavage complex

Previous work has suggested that signaling proteins, and particularly genes in the imd pathway, evolve by positive selection in the *D. simulans* lineage (Schlenke and Begun 2003, Begun and Whitley 2000). While the median P -value for the test of positive selection is marginally lower in genes in the imd pathway than other signaling genes (imd median: 0.1376; other signaling median: 0.2954, $P = 0.050$, Mann-Whitney U, one-tailed), neither signaling genes as a whole nor genes in the imd pathway alone are over-represented among positively selected genes. This discrepancy would be expected if imd pathway genes experience positive selection in only a subset of the species examined. We tested this hypothesis by fitting codon models that allow for lineage-specific variation in ω to test for an acceleration of protein evolution along a particular lineages in the phylogeny (Yang 1998), and by fitting codon models that test explicitly for positive selection that is restricted to particular branches in the phylogeny (Zhang *et al.* 2005).

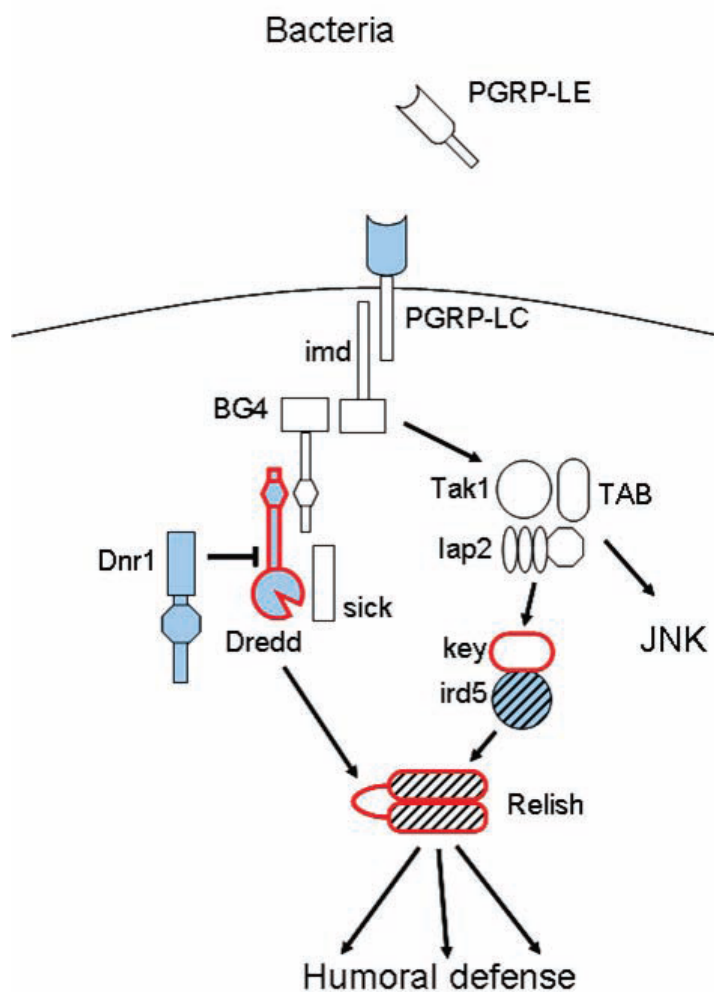


Figure 4.5 Positive selection in the imd pathway. Genes outlined in red have a significantly accelerated ω on the *D. melanogaster* lineage relative to the rest of the tree ($P < 0.01$). Genes filled with diagonal black lines have significant evidence for positive selection along the *D. melanogaster* branch ($P < 0.01$; branch-site test). Genes filled in light blue have significant evidence for positive selection on the entire phylogeny ($P < 0.02$).

A number of genes in the imd pathway (*Relish*, *ird5*, *key*, and *Dredd*) show evidence for accelerated rates of evolution specifically in the *D. melanogaster* lineage at a nominal $P < 0.01$ (Figure 4.5); *Relish* is also accelerated in the ancestral *simulans/sechellia* lineage. Of those four genes, *Relish* and *ird5* also have a subset of codons with $\omega > 1$ specifically in the *melanogaster* lineage; *Dredd*, *Dnr1*, and *ird5* have evidence for positive selection in the entire phylogeny. *BG4* has evidence for positive selection in both the *melanogaster* and *simulans/sechellia* lineages ($P = 0.019$ and $P = 0.048$ respectively), although not in the whole phylogeny. Taken together, these results suggest that a substantial fraction of genes in the imd pathway have experienced positive selection in the *melanogaster* species group, selection that has occurred since the divergence of these species from *D. yakuba* and *D. erecta*, consistent with recent results for *Relish* using a different methodology (Levine and Begun 2007).

Many of the positively selected proteins in the imd pathway are thought to physically interact. Relish is cleaved at a caspase cleavage site located in the spacer region between the N-terminal REL homology domain (encoding the functional transcription factor) and the C-terminal ANK repeat region (encoding an autoinhibitory domain; Stoven *et al.* 2003). Cleavage requires phosphorylation of Relish by the kinase *ird5*, and also requires *Dredd*, a caspase that forms a complex with Relish (Stoven *et al.* 2003). Thus, molecular coevolution may drive positive selection in the spacer region of Relish, the caspase domain of *Dredd*, and the kinase domain of *ird5*. Pooling across all putatively interacting domains, positively selected sites significantly cluster inside the interacting domains (Table 4.2), suggesting that, at least in *D. melanogaster*, the entire complex is evolving by positive selection. The apparent restriction of positive selection in at least some of these genes to the

melanogaster species group suggests that it may stem from a taxon-specific host-pathogen interaction.

Table 4.2. Distribution of positively selected sites among Relish and its interactors

Gene	Domain	Positively Selected		P-value (2)
		Sites (1)	Total Sites	
Relish (3)	Spacer	9	149	0.0033
	Rest	13	815	
ird5 (3)	Kinase	2	300	0.176
	Rest	0	415	
Dredd (4)	Caspase	2	204	0.197
	Rest	0	255	
Total	Interaction			
	Domains	13	653	
	Rest	13	1485	0.018

(1) Any site with a Bayesian posterior probability of positive selection greater than 0.75 is considered a "positively selected site"

(2) Calculated by Fisher's Exact Test

(3) In *D. melanogaster* lineage only

(4) In entire phylogeny

Discussion

A number of the genes we identify as positively selected also evolve adaptively in other organisms, suggesting that widely disparate taxa may yet reveal similarities in the evolution of the immune system, and raising the tantalizing

possibility that certain kinds of immune proteins may generally be involved in host-pathogen ‘arms races.’ The most striking example of such commonalities are the TEP proteins: analysis of fragments of thioester-containing proteins have suggested adaptive evolution in Anopheline mosquitoes (Little and Cobbe 2005) and the cladoceran crustacean *Daphnia* (Little *et al.* 2004), suggesting that the TEP superfamily is commonly the target of positive selection in arthropods and motivating further study in mammals of TEP, α 2-macroglobin, and complement superfamily proteins.

We and others (Begun and Whitley 2000; Schlenke and Begun 2003; Levine and Begun 2007) have also identified *Relish* and its protein interactors as targets of positive selection in *Drosophila*, apparently in only a subset of lineages. In an interesting parallel, positively selected codons have also been detected in the linker, PEST domain, and caspase cleavage site of *Relish* in termites of the genus *Nasutitermes*, with a similar clade-restricted pattern (Bulmer and Crozier 2006), suggesting that *Relish* may commonly be involved in taxon-specific host-pathogen interactions.

Neither our study, nor any previous study (Clark and Wang 1997; Lazzaro and Clark 2003; Jiggins and Kim 2005), has found any evidence for adaptive evolution among AMPs in *Drosophila*. In contrast, AMPs in frogs, termites, and mammals have all be shown to evolve both by rapid gene duplication and positive selection (Tennessen 2005). While we see extensive gene duplication and high rates of gene turnover in *Drosophila*, the lack of positive selection is puzzling. AMPs in *Drosophila* and other insects, in contrast to organisms with adaptive immune systems, serve the role as the primary microbial- and fungal-killing proteins, and may be particularly important in preventing infection by non-coevolving saprophytic organisms, as opposed to more specific pathogens that would be expected to drive rapid

coevolutionary arms races. Furthermore, in the *Drosophila* immune response a large number of different AMPs are induced to high systemic levels after infection. These two factors may lead to stronger selection for speed and efficiency of transcription and translation of AMPs after infection, as opposed to modifications of the protein sequence by positive selection.

The intersection of comparative genomics and molecular evolution provides fertile ground to explore evolution of innate immune pathways along a multispecies phylogeny. While considerable attention has focused on the evolutionary dynamics of components of the adaptive immune system in vertebrates, the results presented here suggest that the *Drosophila* innate immune system experiences similar selective pressures, driven in similar ways by host-pathogen coevolutionary dynamics. Specifically, we find that proteins involved in pathogen recognition, and in particular regions of these proteins that interact with pathogens, experience significantly more positive selection than other components of the innate immune system, reminiscent of classic examples of adaptive evolution in vertebrate immunity. Further work will be needed to assess the generality of the patterns we observe in taxa that have both adaptive and innate immune systems. Nonetheless, the deep genetic resources of *Drosophila* provide a unique opportunity to further understand the functional divergence of innate immune pathways.

Methods

Annotation of immune-related proteins in D. melanogaster

We generated an initial list of immune system proteins in *D. melanogaster* from recent reviews, FlyBase annotations, and the published literature, including any protein with direct molecular evidence for an immune role in *D. melanogaster*, as well as proteins homologous to known immune proteins in *D. melanogaster* or other organisms.

Further details of our functional classifications are described in Sackton *et al.* (2007). All analyses, unless otherwise noted, used this manually curated gene set.

Initial data set and alignments

Our annotation of homologs of immune system genes in non-*melanogaster* species started with predicted GLEANR models and homology clusters derived from the computational analysis described in Clark *et al.* (2007). Briefly, an all-against-all TBLASTN search was run using all annotated *D. melanogaster* proteins and the set of translated consensus gene predictions for each non-*melanogaster* generated by combining several *ab initio* and homology-based gene predictors using GLEAN. The results from this TBLASTN search were combined using a fuzzy reciprocal BLAST algorithm to generate homology clusters, which formed the basis of the primary computational annotation (Clark *et al.* 2007). We then improved these annotations manually. In many cases we were able to extend partial computationally defined gene models (although not always, as occasionally assembly gaps prevented the extension of gene models), merge single models inappropriately split into multiple models, eliminate erroneous paralogy calls introduced by assembly duplications, and find novel homologs not identified by the computational pipeline. In some cases we corrected erroneous frame-shift or nonsense mutations by examining raw sequence traces in the NCBI trace archive.

We derived initial homology assignments from the fuzzy reciprocal BLAST homology clusters described in Clark *et al.* (2007), and refined them by manual annotation. For cases where we judged GLEANR models to be correct, and no paralogs were identified, we used the alignments produced by Clark *et al.* (2007) for all subsequent analyses. In all other cases, we used alignments produced by T-COFFEE and manually edited. We then masked these alignments as described in Clark *et al.* (2007) prior to molecular evolutionary analysis.

Gene family evolution

Many of the genes involved in innate immunity are organized in clusters of related genes having similar function, which can expand or contract in number across the *Drosophila* phylogeny. Phylogenetic hypotheses were used to assign paralogy and orthology within these genes families. Briefly, we started with homology clusters identified by computational algorithms, supplemented by manual TBLASTN searches where necessary. These represent sets of genes that include all orthologs and paralogs of any given *D. melanogaster* gene identifiable in any of the 12 sequenced species. We then assigned these homology clusters to one of three homology classes, based on copy number conservation. Genes that have a single ortholog in all 12 species (*i.e.*, where all pairwise BLAST searches agree on only a single best reciprocal BLAST hit in each species) are assigned to the single-copy orthology class. Genes that have identifiable orthologs and paralogs in both the *Drosophila* and *Sophophora* subgenera (suggesting that these genes were present in the common ancestor), but that are not in the single-copy orthology class, are assigned to the conserved paralogy class. The remaining genes (those that appear to have originated more recently than the common ancestor of the *Drosophila* and *Sophophora* subgenera) are assigned to the lineage-restricted class.

We used birth-death models, which assume that gene families evolve by duplication to create new gene copies with some birth rate and by pseudogenization and loss of existing gene copies with some death rate, to test for variation in rates of gene turnover across families. Using the EM algorithm implemented the CAFE software (De Bie *et al.* 2006), we estimated λ , the rate of gene copy turnover per million years, for each immune system gene family in our dataset, assuming the time to the most recent common ancestor of Drosophilids is 40 million years. Gene families with no copy number variation are assumed to have $\lambda = 0$. CAFE assumes a model of

gene family evolution with a single constant rate for both gene duplication and gene loss that is homogenous across the phylogeny. We also used this analysis to test for non-homogeneity of the birth-death process, although no gene family in our data set rejects non-homogeneity after multiple test correction.

PAML analysis

All PAML analyses were carried out with PAML v3.15 on the *melanogaster* group alignments, described in Clark *et al.* (2007), Sackton *et al.* (2007), and Larracunte *et al.* (2008). For all alignments, we ran PAML model M0, M7, and M8. Model M0 assumes a single ω for each gene, whereas M7 and M8 allow ω to vary among codons in a gene. In general, we use per-gene estimates of ω from M0 unless otherwise noted, and use more complicated models primarily to test for evidence for positive selection. M7 assumes that ω follows a beta(0,1) distribution, with shape parameters estimated by maximum likelihood. M8 makes the same initial assumption, and adds a class of codons with $\omega \geq 1$. Our test for positive selection is a comparison of twice the difference in likelihoods between model M7, which does not allow for positive selection, and M8, which does. We estimated p-values by simulation under the null model, as described in Larracunte *et al.* (2008). We corrected for multiple testing using two different false-discovery-rate approaches, as described in Sackton *et al.* (2007) and Larracunte *et al.* (2008).

For the genes in the *imd* pathway, we also analyzed a series of branch models that allow ω to vary among branches (Yang 1998): one in which the *melanogaster* terminal lineage has one ω , and the rest of the tree has another, and one in which the *simulans* and *sechellia* lineages have one ω , and the rest of the tree has another. These models test for changes in constraint along a particular branch, not positive selection *per se*. In order to explicitly test for positive selection, we used a branch-site model (Zhang *et al.* 2005), which allows four classes of codons: a strictly conserved class (ω

< 1), a class that is conserved on the ‘background’ lineages but under positive selection in the ‘foreground’ lineage of interest, a class that is strictly neutral ($\omega = 1$), and a class that is neutral on the ‘background’ lineages but under positive selection in the ‘foreground’ lineage of interest. When compared to the null model which does not allow positive selection in the foreground lineage, this model is a robust test for positive selection in a particular set of codons on a particular lineage (Zhang *et al.* 2005). We applied this branch-site model to two sets of foreground lineages: the *melanogaster* terminal branch, and the *simulans/sechellia* clade. For both the branch test and the branch-site test, significance was assessed using standard asymptotic assumptions, since both tests are well-behaved (Zhang *et al.* 2005). We used the Bayesian empirical Bayes approach implemented in PAML Model M8 and the PAML branch-site models to estimate the probabilities of positive selection for specific codons (Yang *et al.* 2005).

Statistical analysis

All statistical analysis were carried out in R (version 2.4.1), with the exception of some permutation tests, which we implemented with custom Perl scripts.

REFERENCES

- Agaisse H, Perrimon N (2004) The roles of JAK/STAT signaling in *Drosophila* immune responses. *Immunological Reviews* 198: 72-82.
- Begun DJ, Whitley P (2000) Adaptive evolution of relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics* 154(3): 1231-1238.
- Blandin S, Levashina EA (2004) Thioester-containing proteins and insect immunity. *Molecular Immunology* 40(12): 903-908.
- Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Developmental Cell* 3(5): 711-722.
- Bulet P, Hetru C, Dimarcq JL, Hoffmann D (1999) Antimicrobial peptides in insects; structure and function. *Developmental and Comparative Immunology* 23(4-5): 329-344.
- Bulmer MS, Crozier RH (2006) Variation in positive selection in termite GNBPs and Relish. *Molecular Biology and Evolution* 23(2): 317-326.
- Chang CI, Chelliah Y, Borek D, Mengin-Lecreulx D, Deisenhofer J (2006) Structure of tracheal cytotoxin in complex with a heterodimeric pattern-recognition receptor. *Science* 311(5768): 1761-1764.
- Chang CI, Ihara K, Chelliah Y, Mengin-Lecreulx D, Wakatsuki S *et al.* (2005) Structure of the ectodomain of *Drosophila* peptidoglycan-recognition protein LCa suggests a molecular mechanism for pattern recognition. *Proceedings of the National Academy of Sciences of the United States of America* 102(29): 10279-10284.
- Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S *et al.* (2002) Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298(5591): 159-165.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.
- Clark AG, Wang L (1997) Molecular population genetics of *Drosophila* immune system genes. *Genetics* 147(2): 713-724.

- Daibo S, Kimura MT, Goto SG (2001) Upregulation of genes belonging to the drosomycin family in diapausing adults of *Drosophila triauraria*. *Gene* 278(1-2): 177-184.
- De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22(10): 1269-1271.
- De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *The EMBO Journal* 21(11): 2568-2579.
- Evans CJ, Hartenstein V, Banerjee U (2003) Thicker than blood: conserved mechanisms in *Drosophila* and vertebrate hematopoiesis. *Developmental Cell* 5(5): 673-690.
- Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL *et al.* (2006) Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology* 15(5): 645-656.
- Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* 15(8): 1153-1160.
- Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P *et al.* (2007) Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science*. 317: 1753 - 1756
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335(6186): 167-170.
- Hultmark D (1993) Immune reactions in *Drosophila* and other insects: a model for innate immunity. *Trends in Genetics* 9(5): 178-183.
- Hultmark D (2003) *Drosophila* immunity: paths and patterns. *Current Opinion in Immunology* 15(1): 12-19.
- Jiggins FM, Kim KW (2005) The evolution of antifungal peptides in *Drosophila*. *Genetics* 171(4): 1847-1859.
- Kocks C, Cho JH, Nehme N, Ulvila J, Pearson AM *et al.* (2005) Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in *Drosophila*. *Cell* 123(2): 335-346.

- Kurucz E, Zettervall CJ, Sinka R, Vilmos P, Pivarcsi A *et al.* (2003) Hemese, a hemocyte-specific transmembrane protein, affects the cellular immune response in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 100(5): 2622-2627.
- Kurucz E, Markus R, Zsamboki J, Folkl-Medzihradsky K, Darula Z *et al.* (2007) Nimrod, a putative phagocytosis receptor with EGF repeats in *Drosophila* plasmatocytes. *Current Biology* 17: 649-654
- Larracuenta, AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, *et al.* (2008) Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics* 24: 114-123.
- Lazzaro BP, Clark AG (2003) Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. *Molecular biology and evolution* 20(6): 914-923.
- Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annu Rev Immunol* 25: 697-743.
- Levine MT, Begun DJ (2007) Comparative population genetics of the immunity gene, Relish: is adaptive evolution idiosyncratic? *PLoS ONE* 2: e442.
- Little TJ, Cobbe N (2005) The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Molecular Biology* 14(6): 599-605.
- Little TJ, Colbourne JK, Crease TJ (2004) Molecular evolution of daphnia immunity genes: polymorphism in a gram-negative binding protein gene and an alpha-2-macroglobulin gene. *Journal of Molecular Evolution* 59(4): 498-506.
- Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews* 4(11): 865-875.
- Markow TA, O'Grady PM (2007) *Drosophila* Biology in the Genomic Age. *Genetics*. 177: 1269-1276.
- Meister M, Lagueux M (2003) *Drosophila* blood cells. *Cellular Microbiology* 5(9): 573-580.
- Philips JA, Rubin EJ, Perrimon N (2005) *Drosophila* RNAi screen reveals CD36 family member required for mycobacterial infection. *Science* 309(5738): 1251-1253.

- Ramet M, Pearson A, Manfrulli P, Li X, Koziel H *et al.* (2001) *Drosophila* scavenger receptor CI is a pattern recognition receptor for bacteria. *Immunity* 15(6): 1027-1038.
- Ramos-Onsins S, Aguade M (1998) Molecular evolution of the Cecropin multigene family in *Drosophila*. functional genes vs. pseudogenes. *Genetics* 150(1): 157-171.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D *et al.* (2007) Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics* 39(12): 1461-1468.
- Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system evolution. *Genetics* 164(4): 1471-1480.
- Silverman N, Maniatis T (2001) NF-kappaB signaling pathways in mammalian and insect innate immunity. *Genes & Development* 15(18): 2321-2342.
- Steiner H (2004) Peptidoglycan recognition proteins: on and off switches for innate immunity. *Immunological Reviews* 198: 83-96.
- Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100(16): 9440-9445.
- Stoven S, Silverman N, Junell A, Hedengren-Olcott M, Erturk D *et al.* (2003) Caspase-mediated processing of the *Drosophila* NF-kappaB factor Relish. *Proceedings of the National Academy of Sciences of the United States of America* 100(10): 5991-5996.
- Stroschein-Stevenson SL, Foley E, O'Farrell PH, Johnson AD (2006) Identification of *Drosophila* gene products required for phagocytosis of *Candida albicans*. *PLoS Biology* 4(1): e4.
- Tennessen JA (2005) Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *J Evol Biol* 18(6): 1387-1394.
- Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316(5832): 1738-1743.
- Werner T, Borge-Renberg K, Mellroth P, Steiner H, Hultmark D (2003) Functional diversity of the *Drosophila* PGRP-LC gene cluster in the response to lipopolysaccharide and peptidoglycan. *The Journal of Biological Chemistry* 278(29): 26319-26322.

- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* 13(5): 555-556.
- Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* 15(5): 568-573.
- Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* 22(4): 1107-1118.
- Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155(1): 431-449.
- Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* 22(12): 2472-2479.

CHAPTER 5

COMPARATIVE PROFILING OF THE TRANSCRIPTIONAL RESPONSE TO INFECTION IN TWO SPECIES OF *DROSOPHILIA* BY SHORT-READ cDNA SEQUENCING

Introduction

Host-pathogen interactions are ubiquitous in nature, leading to coevolutionary dynamics that are predicted to drive rapid evolution of the immune system. It is now increasingly clear that this coevolutionary “arms race” leads to increased rates of protein evolution in genes encoding components of the immune system across a large number of taxa (Hughes and Nei 1988; Murphy 1993; Begun and Whitley 2000; Schlenke and Begun 2003; Obbard *et al.* 2006; Tiffin and Moeller 2006; Sackton *et al.* 2007). Recent work in mosquitoes (Christophides *et al.* 2002; Waterhouse *et al.* 2007) and fruit flies (Sackton *et al.* 2007) has suggested that immune system genes may also be unusual in the rate at which new genes are recruited into the system, and existing components of the system turn over by gene duplication or loss. Genes encoding effector proteins (proteins involved in bacterial killing and clearance), and particularly antimicrobial peptides (AMPs), often have lineage-restricted patterns of homology and show very rapid rates of gene turnover within gene families. This is in contrast to genes encoding components of immune-related signaling pathways, which are typically found as single copy orthologs even between distantly related insects (Christophides *et al.* 2002; Evans *et al.* 2006; Waterhouse *et al.* 2007), and have identifiable homologs in mammals (Silverman and Maniatis 2001). Together, these observations that disruption of stoichiometry and other conserved interactions among signaling pathways is usually deleterious, leading to very low tolerance of gene copy

variation among signaling pathways and preservation of single-copy orthologs across deep evolutionary time. Conversely, pathway outputs retain flexibility, with novel effectors easily recruited into the system simply by the acquisition of an immune-responsive promoter element, allowing the possibility of rapid, and perhaps advantageous, proliferation of effector components across evolutionary time.

The genus *Drosophila* provides an ideal system to test this evolutionary network model in more detail. The immune response of *D. melanogaster* has been extensively characterized by more than a decade of careful molecular biology (reviewed in Lemaitre and Hoffmann 2007). These studies have revealed a humoral immune response characterized by the rapid transcriptional upregulation of a wide range of immune-responsive genes in response to bacterial infection (De Gregorio *et al.* 2001; Boutros *et al.* 2002), as well as a cellular response that involves the phagocytosis of foreign microorganisms by circulating macrophage-like cells (Meister 2004). The *D. melanogaster* humoral immune response is initiated when circulating and membrane-bound receptor proteins (members of the PGRP and GNBP gene families) recognize cell wall components (i.e., PGN) of foreign microorganisms (Steiner 2004). This recognition leads to signaling through two primary pathways, the Toll and Imd pathways (De Gregorio *et al.* 2002), which leads to the translocation of the NF- κ B transcription factors DIF, Dorsal, and Relish into the nucleus and the induction of the transcriptional response described above.

Twelve species of *Drosophila* have sequenced genomes, spanning a range of evolutionary distances (Clark *et al.* 2007). Analysis of these genomes has revealed not only signatures of positive selection on immune system genes, but also evidence for rapid turnover of effector genes and acquisition of novel effector components during the evolutionary history of the *Drosophila* genus (Sackton *et al.* 2007). Most strikingly, two multigene families – the Drosomycin antimicrobial peptide family

(Fehlbaum *et al.* 1994) and the Turandot family (Ekengren and Hultmark 2001; Ekengren *et al.* 2001) of induced but otherwise uncharacterized proteins – appear to be evolutionary novelties restricted to the *melanogaster* group and the *melanogaster* and *obscura* groups respectively (Sackton *et al.* 2007). While the exact function of the Turandots is unclear, the Drosomycins are strongly induced by infection in *D. melanogaster*, and play a clear role in the humoral immune response.

The observation of rapid turnover and lineage-restricted effector proteins in *D. melanogaster* raises the question: how general is this pattern across the *Drosophila* genus? If our model is correct, effector proteins should be recruited and lost from the immune system at a relatively high frequency, which implies that novel components of the immune system remain to be discovered in species of *Drosophila* distantly related to *D. melanogaster*. It has long been recognized that highly divergent insect clades often harbor unique antimicrobial peptides: gambicin in mosquitoes (Vizioli *et al.* 2001), lebecin in Bombyx (Hara and Yamakawa 1995), thanatin from the bug Podisus maculiventris (Fehlbaum *et al.* 1996), and many others (reviewed in Bulet *et al.* 1999). However, the evolutionary dynamics of the acquisition of novel effector components in the innate immune system have not been considered previously. In this study, we have characterized the transcriptional response to infection *D. virilis* in an attempt to comprehensively identify immune-induced molecules both with and without homology to *D. melanogaster* immune system components. We also compare the relative induction of members of effector gene families to assess the extent to which rapid turnover in gene content causes divergence at the level of transcription.

D. virilis is a member of the *Drosophila* subgenus that last shared a common ancestor with *D. melanogaster* and the rest of the *Sophophora* subgenus 40 million years ago (Markow and O'Grady 2007). Annotation of the *D. virilis* immune system based on homology to *D. melanogaster* reveals largely identical signaling pathways,

but significant turnover among effector gene families (Sackton *et al.* 2007). Homology based annotation, however, cannot reveal immune components recruited to the *D. virilis* immune response since its divergence with *D. melanogaster*. Short-read sequencing technology provides a method for identifying differentially expressed genes without the limitation of requiring prior annotations, and thus is ideal for detected induced components of the *D. virilis* immune system that lack homologs in *D. melanogaster*. Here, we demonstrate that short-read sequencing of oligo(dT)-primed double-stranded cDNA, prepared as one would for expression microarray hybridization, provides a robust and accurate method to identify differentially expressed regions of the genome. We then use this approach to sequence cDNA from infected and uninfected samples of *D. virilis* to characterize the genes that are induced by infection, and use that sequencing data to annotate novel components of the *D. virilis* immune system.

Aligning sequencing reads to the reference genome

We generated between 4.8 and 5.2 million 36 bp reads from one lane of Solexa sequencing for each of four biological samples: naïve (uninfected) *D. melanogaster* iso-1 (DmelU), 12 hours post-challenge (infected) *D. melanogaster* iso-1 (DmelI), naïve (uninfected) *D. virilis* 15010-1051.87 (DvirU), and 12 hours-post challenge (infected) *D. virilis* 15010-1051.87 (DvirI). Prior to mapping these sequencing reads to the reference genome, we filtered low complexity reads, low quality reads, and repetitive reads (including polyadenylated mRNA tails), resulting in between 1.6 and 2.6 million reads from each of the four samples (Figure 5.1; see Methods for details). To map reads to the reference, we used a combination of Mosaik (a BLAT-like tool optimized for aligning short-read sequencing reads to a reference; A. Quinlan and G. Marth, unpublished) and BLAST (Altschul *et al.* 1990), which allowed us to map

between 71.3% and 83.7% of the reads that passed our filters, representing 1.2 to 1.9 million reads (Figure 5.1; see Methods for details).

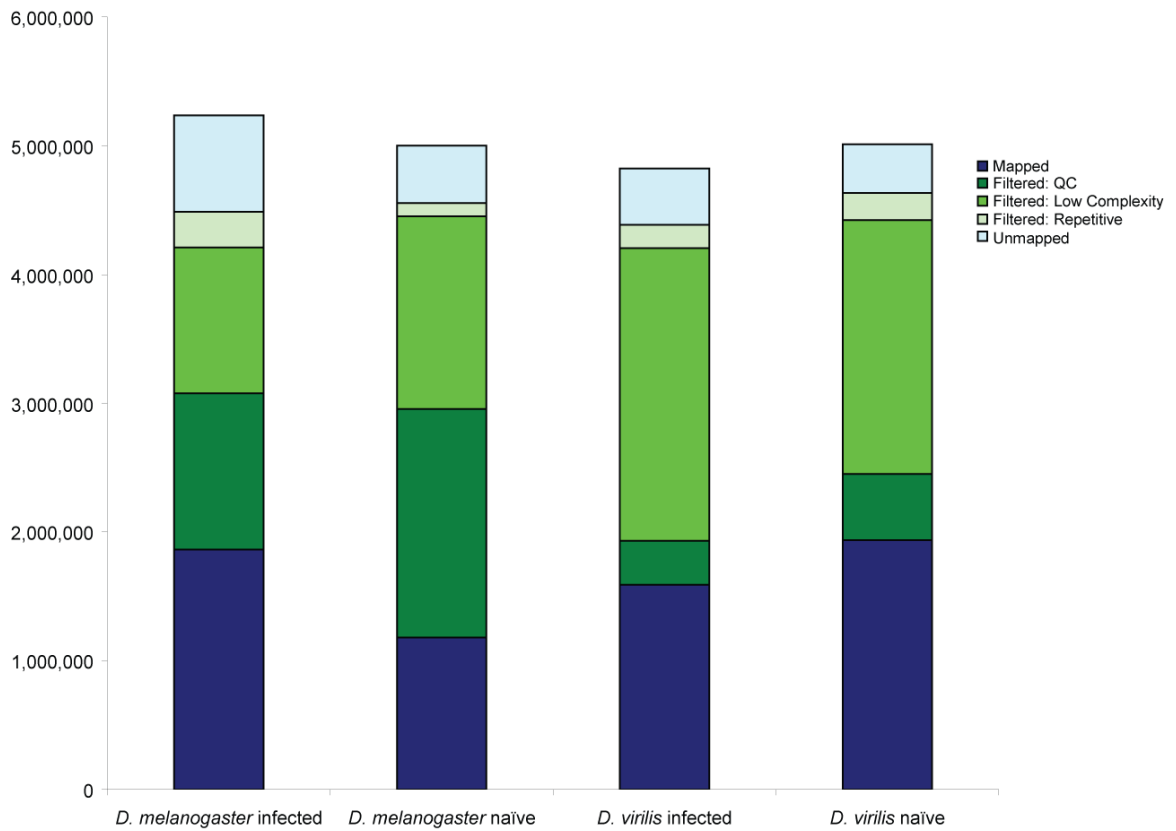


Figure 5.1.

Total number of reads that successfully mapped to the reference genome (dark blue), were filtered prior to mapping (green), and failed to map despite passing our filters (light blue).

Identifying expressed regions regulated by infection

Identifying regions of the genome regulated by infection requires first identifying expressed regions, and then characterizing the between-treatment difference in expression for each region. We defined an expressed region as a segment

of the genome where every base is covered by at least one sequencing read, and the average depth of coverage across both samples combined (infected plus uninfected) is at least 10. Overall, we identify 4,615 expressed regions in *D. melanogaster*; 3,001 of those regions were associated with a total of 2,540 annotated genes. In *D. virilis*, we identified 6,737 expressed regions, of which 584 were associated with 579 genes. The primary reason for the vast discrepancy in the number of expressed regions that could be associated with genes between the two species is the lack of annotated 3' UTRs in *D. virilis*; since our RNA sample was primed from the poly-adenylated tail, our sample has a strong 3' bias. Many of the *D. virilis* regions not overlapping annotated genes are very close to the 3' end of annotated or predicted genes, suggesting that these expressed regions represent unannotated 3' UTR of predicted *D. virilis* genes (see below and Methods for more details).

We then used a Hidden Markov model (HMM) to characterize the transcriptional regulation of each base as one of the following states: strongly induced by infection, induced by infection, not regulated by infection, repressed by infection, or strongly repressed by infection. Each state is defined in the HMM by the binomial probability of observing the number of reads aligned to each base from the infected sample, given the total number of reads that align to each base; because the number of reads that map from the infected and naïve samples is not equal, the binomial probability for the unregulated class is not expected to be 0.50 (Table 5.1; see Methods for details). We define the transcriptional regulation of any given expressed region as the state with the highest weighted sum of assignments to a region (sum of the total coverage of each state across the expressed region). For the 5.2% of expressed regions in *D. virilis* and 13.5% of expressed regions in *D. melanogaster* for which either the weighted sum assigned to the best state was not at least 50% of the total weighted sum, or for which no single base had sufficiently high coverage to be included in our

Table 5.1. Binomial probability of observing x infected reads aligned to a base, out of n total reads aligning to that base, estimated based on the HMM described in the text, and median corrected induction (log2 scale) for each class (see Methods for details).

	<u><i>D. melanogaster</i></u>		<u><i>D. virilis</i></u>	
	Binomial probability	Median induction	Binomial probability	Median induction
Strongly induced	0.9034	5.60	0.9950	107.08
Induced	0.7078	1.68	0.7002	2.44
Not regulated	0.5722	1.10	0.4257	0.99
Repressed	0.4607	0.79	0.3170	0.64
Strongly repressed	0.1575	0.17	0.1766	0.35

HMM, we consider the state as “not determined”. The number of expressed regions assigned to each state are presented in Table 5.2.

Table 5.2. Number of expressed regions assigned to each induction class

	<i>D. melanogaster</i>	<i>D. virilis</i>
Strongly induced	107	33
Induced	747	808
Not regulated	1793	2684
Repressed	1280	2444
Strongly repressed	66	416
Not determined	622	352

Validating D. melanogaster induced regions

Because the transcriptional response to infection has been extremely well characterized in *D. melanogaster* (e.g. De Gregorio *et al.* 2001; Boutros *et al.* 2002; De Gregorio *et al.* 2002; Apidianakis *et al.* 2005; Irving *et al.* 2005), we can validate our approach by comparison to previous studies. We compiled data from four

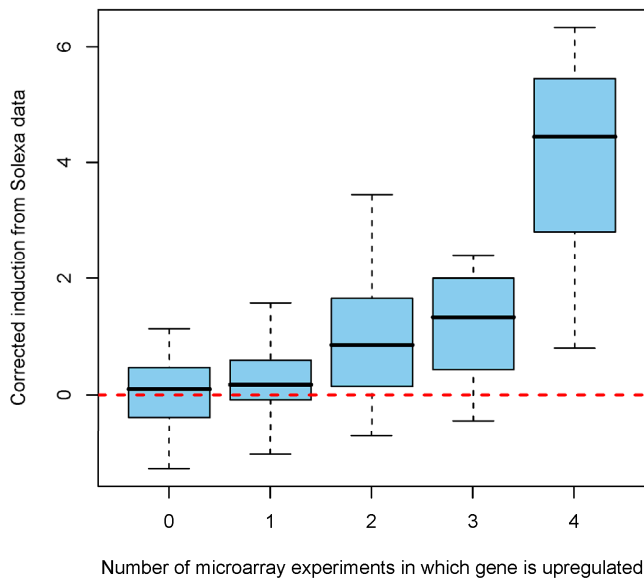


Figure 5.2.

Boxplot of corrected induction measured by short-read sequencing for genes detected as upregulated in 0, 1, 2, 3, or 4 previous microarray studies that used a similar infection design (see text for details). Induction is plotted on a log2 scale.

microarray experiments published between 2001 and 2005 (De Gregorio *et al.* 2001; Boutros *et al.* 2002; Apidianakis *et al.* 2005; Irving *et al.* 2005) that compared gene expression in infected and naïve *D. melanogaster* and characterized genes as up-

regulated or down-regulated. Based on the definitions from each study, we count how many times a gene was defined as ‘up-regulated’ or ‘down-regulated’ across the four studies. There are 294 genes that are both present in our list of expressed regions and significantly regulated by infection on at least one of the four microarrays; those that are upregulated in multiple microarrays are substantially more likely to be strongly induced in our data (Figure 5.2). Furthermore, genes induced in more microarrays are more likely to be assigned to an induced state by our HMM (Figure 5.3).

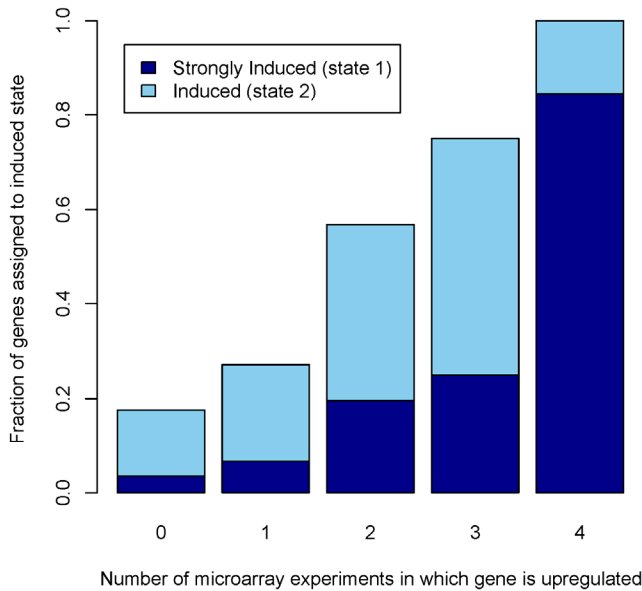


Figure 5.3.

The fraction of genes in each microarray induction category that are assigned to an induced state based on our Solexa sequencing data.

Finally, we used the previously published microarray most similar to our experiment (De Gregorio *et al.* 2001), in which *D. melanogaster* was infected by septic injury with a mixed bacterial culture, assayed at 12 hours post infection, to

make a quantitative comparison of induction. Despite differences in the line (Oregon R vs. iso-1) and sex (male vs. female) of the flies, and the species and pathogenicity of bacteria used (non-pathogenic *E. coli* and *M. luteus* mixture vs. pathogenic *S. marcescens* and *E. faecalis* mixture), we still find a significant correlation between induction measured by microarray in DeGregorio *et al.* (2001) and induction measured by our method ($r = 0.3225$, $P < 2.2 \times 10^{-16}$). Taken together, these results demonstrate considerable consistency between induction measured by short-read sequencing of oligo(dT) primed, double-stranded cDNA and induction measured by traditional microarray methods. At least for identifying strongly induced genes, short read sequencing approaches appear to be robust and accurate, suggesting that this approach may prove to be a simple and cost-effective way to identify differentially regulated genes in poorly annotated genomes in response to any number of treatments of biological interest.

The transcriptional response to infection in D. virilis and D. melanogaster

We identified 841 expressed regions that appear to be induced by infection in *D. virilis*. Because of the 3' bias in our cDNA preparation, the relatively low coverage we sequenced to, and the lack of annotation of 3' UTR sequence in the *D. virilis* genome, only about 5% of these induced regions overlap with an annotated exon. In order to attempt to associate a greater percentage of induced regions with genes, we analyzed the genomic region in more detail for these 841 regions, and preliminarily assigned expressed regions to annotated gene models if they were less than 500 bp from the 3' end of the nearest gene model, and more than 1 kb from the 3' end of all other genes models (see Materials and Methods for details). We eliminated from further analysis induced regions (but not strongly induced regions) located on minor scaffolds (< 1 megabase), leaving a total of 199 candidate induced regions in *D. virilis*, 101 of which can be preliminarily associated with 95 *D. virilis* genes.

In order to understand the similarities between the *D. melanogaster* and *D. virilis* immune responses, we focused on 490 genes in *D. melanogaster* and 95 in *D. virilis* associated with induced regions. We used three approaches to identify orthologs and paralogs of these gene models. First, for any gene model included in the manual homology annotation of immune system genes (Sackton *et al.* 2007), we used the homology and orthology assignments from that work. For the remaining genes, we used homologs annotated by FlyBase. If FlyBase reported no homolog, we verified the absence of homologs by reference to the homology assignments generated by the *Drosophila* 12 Genomes Consortium (Clark *et al.* 2007).

Of the 490 induced genes in *D. melanogaster*, 19 have no identifiable homologs in *D. virilis*, 444 have homologs in *D. virilis*, and 27 have ambiguous homology. Genes associated with expressed regions assigned to state 1 (highly induced) are significantly more likely to lack homologs to *D. virilis* (Odds ratio = 5.22, Fisher's Exact Test *P*-value = 0.001). As highly induced genes are much more likely to represent effectors, this result is expected based on the analysis of Sackton *et al.* (2007), which showed that effector proteins are much more likely to have lineage restricted patterns of homology than immune system genes as a whole.

This same pattern seems to hold in *D. virilis*, suggesting that the high rate of turnover in effector proteins may be quite general. In *D. virilis*, the 95 gene models associated with induced regions include 8 with no identifiable homologs in *D. melanogaster* and 87 with homologs in *D. melanogaster*. Like in *D. melanogaster*, genes associated with expressed regions assigned to state 1 are more likely to lack homologs, although this pattern is not significant (Odds ratio = 3.08, Fisher's Exact Test *P*-value = 0.15).

The putatively induced genes that have identifiable homologs between species reveal broad similarities in the transcriptional response to infection between *D.*

melanogaster and *D. virilis*. As expected based on our comparison of the *D. melanogaster* induced genes to previous microarray studies, most of the highly induced genes are antimicrobial peptides, Turandots, and other immune-induced peptides such as the IMs (*Drosophila* immune molecule, Uttenweiler-Joseph *et al.* 1998). We also see other immune genes such as PGRP-SB1, Transferrin 1, and TepII strongly induced after infection in *D. melanogaster*. In *D. virilis*, homologs of many of these genes are also strongly induced. The genes associated with expressed regions assigned to state 1 in *D. virilis* encode attacins, cecropins, metchnikowin, diptericins, PGRP-SB1, and a protein with homology to IM1. In the next section, we focus on the most strongly induced category of genes in both species, the AMPs.

Differences in the members of AMP families induced after infection

Despite the overall similarity of the transcriptional response to infection in *D. virilis* and *D. melanogaster*, notable differences exist in the pattern of induction of members of AMP gene families. The *D. melanogaster* genome encodes 20 antimicrobial peptides that are members of seven gene families. These peptides can be broadly grouped into three categories: cysteine-rich peptides characterized by pairs of disulfide bonds (defensin: Def; and drosomycins: Drs, Drs-1, Dro2, Dro3, Dro4, Dro5, Dro6), peptides with an amphiphilic α -helical conformation (cecropins: CecA1, CecA2, CecB, CecC), and proline or glycine rich peptides (attacins: AttA, AttB, AttC, AttD; drocosin: Dro; metchnikowin: Mtk, and diptericins: Dpt and DptB). Five of these families – diptericins, cecropins, attacins, metchnikowin, and defensin – have homologs in *D. virilis*, encoding a total of 15 known antimicrobial peptides; drocosin and drosomycins are absent from the entire *Drosophila* subgenus (Sackton *et al.* 2007).

Most of these AMPs are strongly induced after infection (Table 5.3 and Table 5.4). In *D. virilis*, 10 of the 15 AMPs are highly induced (state 1) after infection; the

Table 5.3. Induction of antimicrobial peptides in *D. melanogaster*.

AMP family	<i>D. melanogaster</i> gene	HMM state	Induction	Average Coverage (Naïve)	Average Coverage (Infected)
attacins	AttA	strongly induced	21.59	1.33	38.37
	AttB	strongly induced	23.23	2.26	70.20
	AttC	strongly induced	35.90	0.96	46.29
	AttD	ND ¹			
cecropins	CecA1	strongly induced	27.33	1.18	43.20
	CecA2	strongly induced	43.80	0.68	39.59
	CecB	ND			
	CecC	ND			
defensins	Def	moderately induced	3.35	7.68	34.40
diptericins	Dpt	strongly induced	80.26	2.99	320.51
	DptB	strongly induced	Inf	0.00	30.37
metchnikowin	Mtk	strongly induced	23.47	7.94	249.23
drocosin	Dro	strongly induced	17.07	28.03	640.13
drosomycins	Drs	strongly induced	6.99	19.28	180.39
	Dro2	ND			
	Dro3	ND			
	Dro4	not regulated	0.95	12.74	10.00
	Dro5	ND			
	Dro6	ND			
	Drs-l	ND			

¹Not Detected in our sample.

Table 5.4. Induction of antimicrobial peptides in *D. virilis*.

AMP family	<i>D. virilis</i> gene	<i>D. melanogaster</i> homolog	HMM state	Induction	Average Coverage (Naïve)	Average Coverage (Infected)
attacins	dvir_GLEANR_6000	AttA/AttB	strongly induced	Inf	0.00	53.43
	dvir_GLEANR_6001	AttA/AttB	strongly induced	329.41	0.31	76.45
	dvir_GLEANR_6553	AttC	ND ¹			
	dvir_GLEANR_8042	AttD	ND			
cecropins	dvir_GLEANR_10332	CecA1/CecA2/CecB/CecC	strongly induced	94.93	2.73	191.95
	dvir_GLEANR_10659	CecA1/CecA2/CecB/CecC	strongly induced	52.34	1.20	46.65
	dvir_GLEANR_10661	CecA1/CecA2/CecB/CecC	strongly induced	Inf	0.00	35.45
	dvir_GLEANR_10660	CecA1/CecA2/CecB/CecC	ND			
	dvir_GLEANR_10662	CecA1/CecA2/CecB/CecC	ND			
defensins	dvir_GLEANR_7763	Def	strongly induced	147.31	0.78	85.64
	dvir_GLEANR_6510	Def	ND			
diptericins	dvir_GLEANR_5386	Dpt	strongly induced	326.80	0.88	212.92
	dvir_GLEANR_5385	Dpt	strongly induced	3218.05	1.81	4322.99
	dvir_GLEANR_5387	DptB	strongly induced	317.73	0.89	209.23
metchnikowin	dvir_GLEANR_7753	Mtk	strongly induced	58.15	9.29	400.44

¹Not Detected in our sample.

remaining 5 are not expressed in our sample. In *D. melanogaster*, 10 of the 20 AMPs are highly induced, an 11th is moderately induced (state 2), and a 12th appears to be very weakly expressed, but not induced (state 3). Among the homologous AMP families, the dipterocins are the most strongly induced in both species: marginally so in *D. melanogaster* (Table 5.3), and strikingly so in *D. virilis* (Table 5.4). Furthermore, in both species dipterocins represent the largest fraction of total AMP transcription in the infected sample (as measured by average coverage of the infected sample), and dipterocin expression is dominated by a single paralog (Figure 5.4); when non-homologous AMP families are included, Dro dominates overall infected transcription in *D. melanogaster*. The extent to which one AMP predominates is strikingly different between species: in *D. melanogaster*, Dro, Mtk, Drs and Dpt are all transcribed at high levels in the infected sample, whereas no other AMP is transcribed at nearly the level of Dpt in *D. virilis* (Figure 5.4). There are also differences in the relative transcription level of paralogs within the cecropin, attacin, and dipterocin AMP families. In the *D. virilis* sample, one member of each family tends to dominate transcription (Figure 5.4), whereas in *D. melanogaster* the relative transcription of paralogs within a AMP family is less skewed (with the exception of Dpt).

On a broader scale, in both species the proline- and glycine- rich peptides represent most of the total AMP transcription (*D. virilis*: 93.6%, *D. melanogaster*: 81.9% of the total coverage across all AMPs, normalized for length). Again, though, *D. melanogaster* appears to transcribe a broader spectrum of antimicrobial peptides in response to our challenge, with a substantial fraction of the total transcription of AMPs in *D. melanogaster* associated with cysteine-rich (Drs/Def; 13.5%) AMPs. This analysis of course excludes any uncharacterized AMPs in *D. virilis*. However, as discussed below, the most promising candidates for novel *D. virilis* AMPs appear to be in the glycine- and proline- rich family, suggesting that *D. virilis* may in fact

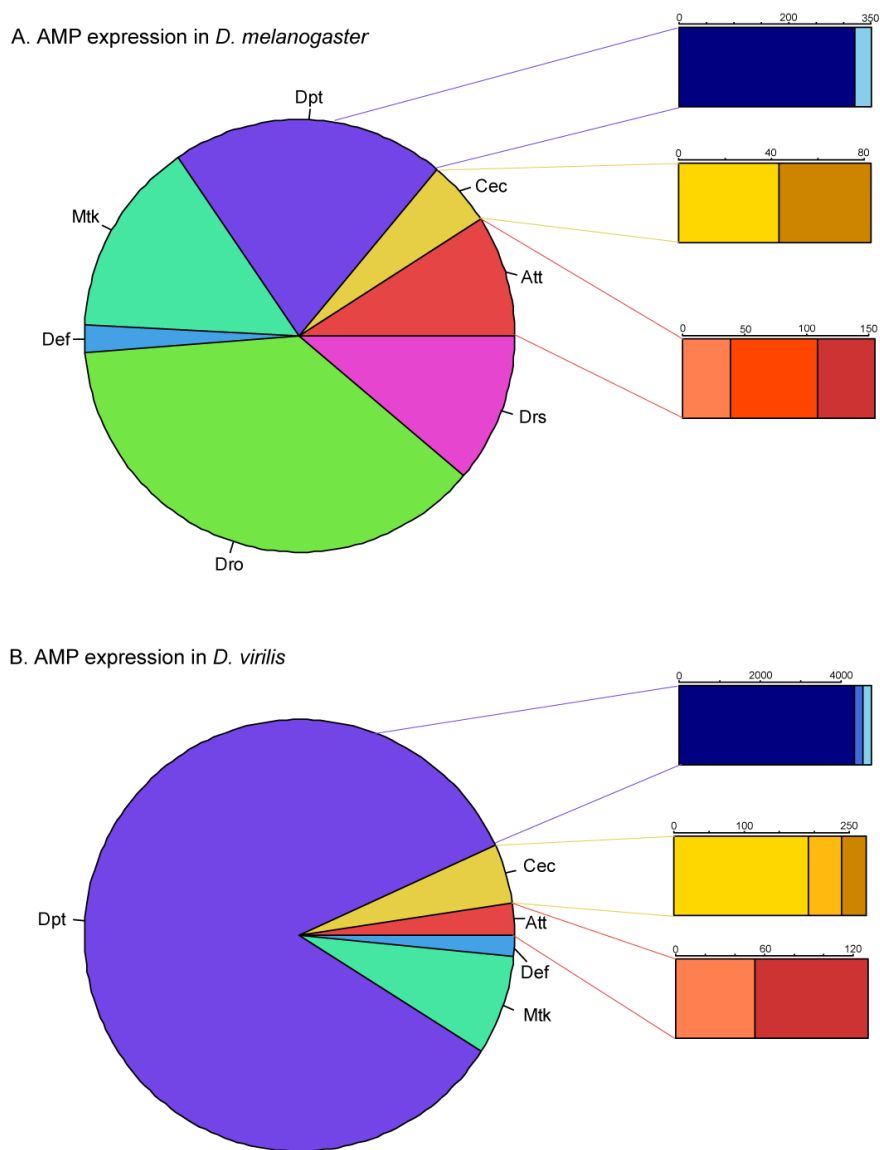


Figure 5.4.

Relative expression of antimicrobial peptides in A) *D. melanogaster* and B) *D. virilis* after infection. Pie charts show the relative expression of AMP families in the infected state. Stacked bar charts show the relative expression of individual AMP genes within families.

produce a narrower range of AMP types. *D. virilis* and *D. melanogaster* differ at a number of ecological traits, any of which could potentially lead to different selective pressures for the diversity of AMPs produced: *D. melanogaster* is tropical, *D. virilis* is Holarctic; *D. melanogaster* breeds on a wide range of substrates, typically rotting fruit, *D. virilis* breeds on sap fluxes (Markow and O'Grady 2007). While it is tempting to speculate, fuller understanding of the diversity of *D. virilis* AMPs, and the persistence of differences in transcription in response to multiple challenges, will be needed before the hypothesis that *D. virilis* produces a narrower and less diverse range of AMPs after infection can be established.

Novel components of the D. virilis immune system

As noted above, a number of *D. virilis* gene models associated with induced regions do not have identifiable homologs in *D. melanogaster* or other species of the *melanogaster* species group. In this section, we discuss the eight *D. virilis* induced GLEANR models with no identifiable homologs outside of the *Drosophila* subgenus, plus an additional pair of strongly induced expressed regions which we propose associate with a novel family of small peptides in *Drosophila* (Table 5.5). Broadly speaking, these ten gene models fall into three classes: those encoding predicted proteins that lack a signal peptide; those encoding predicted proteins that are short, secreted, negatively charged, and appear to be distantly related to the IM proteins in *D. melanogaster*; and those encoding short, positively charged, secreted proteins that are often proline or glycine rich and may represent novel AMPs. Predicted proteins were considered secreted if SignalP (Bendtsen *et al.* 2004) predicted a signal peptide.

Gene models lacking a signal peptide: Two putatively induced *D. virilis* gene models lack a signal peptide. One, dvir_GLEANR_13841 is a short protein (155 aa) that is moderately induced (corrected induction 2.04, assigned to state 2). We identify

Table 5.5. Genes associated with induced regions in *D. virilis* that lack homologs in *D. melanogaster*.

<i>D. virilis</i> gene model	Induction State	Induction	Signal Peptide?	Size (kD)	Net Charge
GF_DGIL_SNO_29059273/ GF_DGIL_SNO_29059274*	1	716.86	yes	4.13	+1
dvir_GLEANR_6300	2	1.55	yes	4.1	+1
dvir_GLEANR_7739	1	Inf	yes	6.13	+9
dvir_GLEANR_5464	1	61.95	yes	4.71	+4
dvir_GLEANR_3774	2	2.89	yes	4.57	+3
dvir_GLEANR_345	1	26.91	yes	2.23	-2
dvir_GLEANR_5361	2	3.14	yes	2.22	-1
dvir_GLEANR_15023	2	2.22	no		
dvir_GLEANR_13841	2	2.04	no		
dvir_GLEANR_3200	2	1.96	no		

*These gene models are paralogs

putative homologs in *D. grimshawi* and *D. willistoni*, but not *D. mojavensis*. The second, dvir_GLEANR_15023, is somewhat longer, 280 amino acids, but is highly repetitive, consisting of 20 repeats of a 12-17 amino acid motif. The repetitive nature of this predicted gene makes identifying putative homologs difficult; we fail to detect any via BLAST, and no homologs in any species are reported in *Drosophila* 12 Genomes Consortium (Clark *et al.* 2007). This gene model is flagged as potentially representing a repeat-contaminated gene model (not surprisingly), suggesting that this result may be artifactual. Furthermore, the expressed region we identify does not overlap the predicted gene, but rather associates with it in a manner consistent with the expressed region representing a 3' UTR of the gene model. While these two gene models could potentially represent novel components of the *D. virilis* immune system, we do not believe they represent particularly strong candidates.

Secreted, IM-like peptides: Two gene models in *D. virilis* that are putatively associated with an induced region, dvir_GLEANR_345 and dvir_GLEANR_5361, have strong evidence for a signal peptide, and are short (less than 50 amino acids), suggesting the possibility that these are novel effector proteins. However, they are unlikely to be antimicrobial peptides. Almost all antimicrobial peptides have a net

positive charge; the predicted proteins encoded by these two GLEANR models both have a negative net charge. The proteins that they appear to have distant homology to are the IM proteins of *D. melanogaster*. These are a family of short, strongly induced peptides of unknown function.

Putative novel AMPs: The remaining three GLEANR models, plus the two non-GLEANR gene models (which are paralogs) that appear to be strongly induced by infection, are all secreted peptides with predicted molecular weights between 4 and 6 kD and predicted positive charge at physiological pHs (Table 5.). All three have homologs in *D. mojavensis* and *D. grimshawi*, but no other species, suggesting they are restricted to the *Drosophila* subgenus. In addition dvir_GLEANR_3774 and the unannotated gene are both ~18% proline, suggesting that they might be similar in function to the proline-rich family of AMPs, which include abaecins and apidaecins from bees, as well as Mtk and Dro from *Drosophila*.

Additional evidence will be needed to verify that these putative proteins are secreted into the hemolymph after infection and have an antimicrobial role. Nonetheless these data, combined with the substantial fraction of induced regions that cannot be associated with any annotated gene, suggest quite strongly that *D. virilis*, like *D. melanogaster*, has acquired novel immune components, particularly in the downstream components of the pathway.

Conclusions

In this study, we used short read cDNA sequencing to characterize the transcriptional response to infection in *D. virilis* (and in *D. melanogaster*, for comparison). We show that even a relatively small number of sequencing reads (1 lane per sample, about 5 million reads before filtering and about 1.2 million mapped reads) can produce reliable estimates of induction, at least for strongly induced genes. By

comparing the relative induction of AMP gene families in *D. melanogaster* and *D. virilis*, we show that significant differences in the relative induction of different peptides exist between species, indicating that the rapid turnover in gene family copy number may have significant consequences on the dynamics of immune system induction. Finally, we show that some predicted *D. virilis* genes that lack homologs to *D. melanogaster*, share characteristics with the proline-rich AMP superfamily, suggesting that *D. virilis* likely possesses lineage-restricted immune system components, and that the pattern we observe in *D. melanogaster* is general. Taken together, these results support a “bow-tie” model of immune system evolution, suggesting that novel downstream components of the immune system can be rapidly integrated of relatively short time scales. The adaptive potential of gene gain and loss should not be overlooked in the evolutionary dynamics of host immune systems.

Methods

Biological samples: The *Drosophila* stocks used in this experiment were the sequenced strains of *D. melanogaster* (iso-1) and *D. virilis* (15010-1051.87). Flies were maintained in bottle cultures on a rich dextrose medium at 25° and in 12hr:12hr light/dark for the duration of the experiment. We infected 50 females of each species with a mixed bacterial culture of *Serratia marcescens* and *Enterococcus faecalis* by pricking the thorax with a 0.1-mm dissecting pin (Fine Science Tools, Foster City, CA) dipped in bacterial culture, as previously described (Lazzaro *et al.* 2004). At 12 hours after infection, infected flies and a sample of 50 naïve flies were frozen in liquid nitrogen. We extracted total RNA from frozen flies using standard protocols (Trizol). After extraction, total RNA was treated with DNase (xxxx) to remove potential genomic DNA contamination, according to the manufacturer’s protocols. We

synthesized first strand cDNA using oligio-d(T) primers, and then synthesized second strand cDNA, according to standard protocols. Solexa sequencing was done by the Cornell BRC.

Aligning reads to the reference genome: Prior to mapping reads to the reference genome, we filtered low quality, low complexity, and repetitive reads. We first removed any read with fewer than 24 bases with a Phred quality score (Q) greater than 20; this is our 'low quality' filter. Next, we removed any read with low nucleotide complexity (80%+ of the sequence composed of only 2 bases) or repetitive elements (more than half the sequence composed of dinucleotide or trinucleotide repeats). Finally, we removed any reads with a mononucleotide run greater than 24 base pairs.

After filtering, we did an initial round of mapping to the repeat-masked *D. virilis* or *D. melanogaster* reference genome with Mosaik, a software program written by the Gabor Marth lab (A. Quinlan and G. Marth, unpublished) that uses a BLAT-like approach. The program hashes the genome into unique n-mers (where n, the hash size, can be specified by the user; we used 17 bp), which it uses as seeds to align the sequencing reads to the reference. We required all matches to align for at least 91% of the read length (33 of 36 bp) and have no more than 3 mismatches.

To supplement the mapping from this initial Mosaik run, we took two approaches. First, we noticed that some reads fail to map because low quality ends or partial polyA sequence cause them to fail to pass our alignment length filter. In order to get around this, we trimmed up to 10 bp from the end of any read where average quality across a 5 or 10 bp segment was less than Phred Q20. We also trimmed any mononucleotide run from the end of a read. After trimming, we rejected any read that was shorter than 20 bp, or that failed to pass our QC filters (which we reran on the

trimmed sequence). The remaining reads were then rerun in Mosaik, using the same parameters described above.

Finally, we attempted to map the remaining sequences using BLAST. Any reads that passed all our quality control filters, but could not be mapped using Mosaik even after end-trimming, were run through a BLAST pipeline: we used `blastn` with a word size of 7 and an E-value cutoff of 1×10^{-6} , and considered any read mapped if either 1) there was only a single BLAST hit to the reference genome, or 2) there were fewer than 10 hits, but the best hit aligned over at least 90% of the read length and had a lower E-value than the next best hit. Any read with over 10 hits was considered repetitive and not mapped.

After mapping, we combined the output from both Mosaik runs and the BLAST pipeline to produce a single file for each contig containing the depth of coverage at each base in the genome (using the program `ace2dep`, from the Marth lab, to convert Mosaik output into depth, and custom perl scripts to convert BLAST output into depth information). The depth of coverage at each base along a scaffold was then the input to our pipeline to identify expressed regions of the genome regulated by infection.

Identifying regions regulated by infection: We first defined an expressed region as any contiguous stretch of DNA along a scaffold where the minimum coverage of the combined infected and uninfected samples at any one base is 1, and the average combined coverage across the region is at least 10. Based on this definition, we identified 4615 expressed regions in *D. melanogaster* and 6737 in *D. virilis*. The median length of an expressed region in *D. melanogaster* is 237 bp (Figure 5.5a), compared to a median length in *D. virilis* of 104 bp (Figure 5.5b). Approximately the same number of reads map to *D. virilis* and *D. melanogaster*

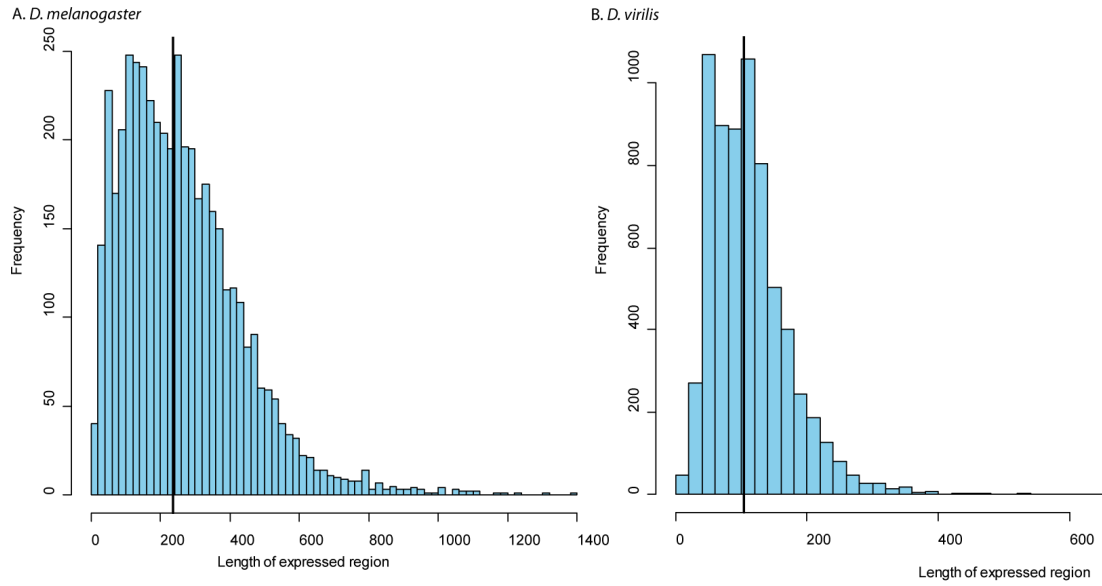


Figure 5.5

Distribution of lengths of expressed regions in A) *D. melanogaster* and B) *D. virilis*.

Solid black lines show the median of each distribution.

(Figure 5.1), so it is unclear why we identify more regions that are on average shorter in *D. virilis*.

To determine the extent to which each expressed region responds to infection, we developed a Hidden Markov Model, with five hidden states representing the degree of induction (highly induced, moderately induced, unchanged, moderately repressed, highly repressed), where the emission probability for each state is the binomial probability of observing X infected coverage given N total coverage at each base pair, and the observed data is the coverage of infected reads at each base. We used the HiddenMarkov package in R to optimize our HMM using the Baum-Welch algorithm, and to calculate the most probable set of states using the Viterbi algorithm. The optimized emission probabilities for each state are given in Table 5.1. Before running the HMM, we removed sites with less than 10x coverage pooled across samples, as there is very little power to distinguish between states with so few reads. In order to

increase the number of expressed regions for which all sites are assigned to the same state, we tuned the transition probabilities by increasing the probability of remaining in the same state, and decreasing the probabilities of transitioning between states proportionally, so that the highest probability in the matrix was equal to 0.999. Empirically, this tuning appears to increase the consistency of our results, with fewer expressed regions being assigned to multiple states.

To determine the most likely state for any given expressed region, we weighted the Viterbi estimate of the state of each base in an expressed region by the summed coverage of that base. If one state had a majority of this measure, we assigned the expressed region to that state; otherwise, we considered the state ambiguous and assigned that expressed region to the “not determined” category (Table 5.2).

Associating expressed regions with genes: We used two methods to associate expressed regions with annotated genes. First, we used an automated first-pass method, where we simply asked whether any base in an expressed region also falls into an annotated exon, based on the *D. melanogaster* release 5.7 annotations and the *D. virilis* release 1.1 annotations available as GFF files from FlyBase. However, given the 3' bias inherent in oligio(dT) primed cDNAs, plus the relatively low coverage that we sequenced to, our identified expressed regions are short (Figure 5.5). In *D. melanogaster* this does not pose much of a challenge, as the genome annotation is quite mature and includes fully annotated 3' UTRs. In *D. virilis*, however, 3' UTRs are generally not annotated, leading to any expressed region that falls entirely in a UTR failing to be associated with any gene. Given the length distribution of 3' UTRs in *D. melanogaster* (Figure 5.6), we expect a substantial fraction of our expressed regions in *D. virilis* to suffer from this problem.

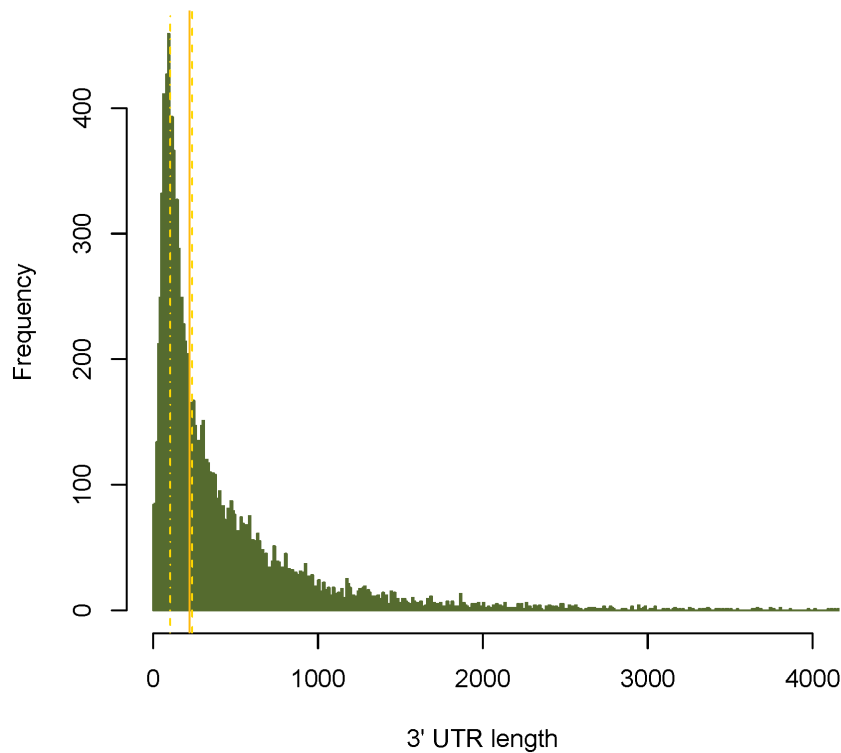


Figure 5.6.

Distribution of 3' UTR length in *D. melanogaster*. The solid gold line is the median of the distribution; the dashed line is the median length of expressed regions in *D. melanogaster*; and the dot-dashed line is the median length of expressed regions in *D. virilis*.

As a partial remedy, we analyzed the 841 induced genomic regions in *D. virilis* in more detail. For these 841 regions, we extracted the nearest 3' end of a gene to the start of the expressed region on the positive strand and the nearest 3' end of a gene to the end of the expressed region on the negative strand. We consider one of these expressed regions to be putatively associated with a annotated gene in *D. virilis* if the smaller distance was less than 500 bp and the larger distance was greater than 1 kb, or if the smaller distance was less than 200 bp and the larger distance was greater than 500 bp. Regions more distant that 500 bp from any other gene were declared

“putatively unassociated,” and the remaining regions were declared “ambiguous.” The number of regions assigned to each class is listed in Table 5.6.

Table 5.6. Association of induced expressed regions with gene models in *D. virilis*.

	Strongly Induced (state 1)	Induced (state 2)
Overlap GLEANR model	10 (30.3%)	34 (4.21%)
Associate with GLEANR model	11 (33.3%)	52 (6.44%)
Ambiguous	1 (3.03%)	31 (3.84%)
Not associated with annotated GLEANR model	11 (33.3%)	691 (85.5%)
Total Regions	33	808

In order to understand why there is a large difference in the fraction of expressed regions that can be associated with annotated genes between the two induced classes in *D. virilis*, we divided the *D. virilis* scaffolds into “major” scaffolds (the 23 scaffolds with at least 1 megabase of sequence, which represent 77% of the total *D. virilis* sequence) and the remaining “minor” scaffolds. Expressed regions assigned to state 1 are much more likely to be on a major scaffold than expressed regions assigned to state 2 (Odds ratio = 14.3, Fisher’s Exact Test P -value = 4.85×10^{-12}). As expected, regions on minor scaffolds, irrespective of class, are much more likely to fail to be associated with an annotated gene (Odds ratio = 144.6, Fisher’s Exact Test P -value $< 2.2 \times 10^{-16}$). However, the difference between minor and major scaffolds does not seem to fully explain the difference between state 1 and state 2, as even when restricted to just the major scaffolds expressed regions assigned to state 1 are more likely to be associated with genes (Odds ratio = 4.4, Fisher’s Exact Test P -value = 0.0027). It could be that highly induced genes are more likely to have homologs in *D. melanogaster*, increasing the probability that those genes would be annotated in *D. virilis*. However, among just the regions that are associated with

genes, it is actually state 2 that is more likely to have homologs in *D. melanogaster* (based on FRB calls; Odds ratio = 3.60, Fisher's Exact Test P -value = 0.0176).

Because of the difficulties in annotating genes on minor scaffolds, we have limited our primary analysis to the 33 expressed regions in state 1, plus the 166 expressed regions in state 2 that are on major scaffolds: this sample of 199 expressed regions includes 101 that can be associated with an annotated gene, as described above, and 98 that cannot.

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3): 403-410.
- Apidianakis Y, Mindrinos MN, Xiao W, Lau GW, Baldini RL *et al.* (2005) Profiling early infection responses: *Pseudomonas aeruginosa* eludes host defenses by suppressing antimicrobial peptide gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 102(7): 2573-2578.
- Begun DJ, Whitley P (2000) Adaptive evolution of relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics* 154(3): 1231-1238.
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340(4): 783-795.
- Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Developmental Cell* 3(5): 711-722.
- Bulet P, Hetru C, Dimarcq JL, Hoffmann D (1999) Antimicrobial peptides in insects; structure and function. *Developmental and Comparative Immunology* 23(4-5): 329-344.
- Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S *et al.* (2002) Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298(5591): 159-165.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.
- De Gregorio E, Spellman PT, Rubin GM, Lemaitre B (2001) Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 98(22): 12590-12595.
- De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *The EMBO Journal* 21(11): 2568-2579.
- Ekengren S, Hultmark D (2001) A family of Turandot-related genes in the humoral stress response of *Drosophila*. *Biochemical and Biophysical Research Communications* 284(4): 998-1003.

- Ekengren S, Tryselius Y, Dushay MS, Liu G, Steiner H *et al.* (2001) A humoral stress response in *Drosophila*. *Current Biology* 11(9): 714-718.
- Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL *et al.* (2006) Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology* 15(5): 645-656.
- Fehlbaum P, Bulet P, Michaut L, Lagueux M, Broekaert WF *et al.* (1994) Insect immunity. Septic injury of *Drosophila* induces the synthesis of a potent antifungal peptide with sequence homology to plant antifungal peptides. *The Journal of Biological Chemistry* 269(52): 33159-33163.
- Fehlbaum P, Bulet P, Chernysh S, Briand JP, Roussel JP *et al.* (1996) Structure-activity analysis of thanatin, a 21-residue inducible insect defense peptide with sequence homology to frog skin antimicrobial peptides. *Proceedings of the National Academy of Sciences of the United States of America* 93(3): 1221-1225.
- Hara S, Yamakawa M (1995) Moricin, a novel type of antibacterial peptide isolated from the silkworm, *Bombyx mori*. *The Journal of Biological Chemistry* 270(50): 29923-29927.
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335(6186): 167-170.
- Irving P, Ubeda JM, Doucet D, Troxler L, Lagueux M *et al.* (2005) New insights into *Drosophila* larval haemocyte functions through genome-wide analysis. *Cellular Microbiology* 7(3): 335-350.
- Lazzaro BP, Scurman BK, Clark AG (2004) Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. *Science* 303(5665): 1873-1876.
- Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annu Rev Immunol* 25: 697-743.
- Markow TA, O'Grady PM (2007) *Drosophila* biology in the genomic age. *Genetics* 177(3): 1269-1276.
- Meister M (2004) Blood cells of *Drosophila*: cell lineages and role in host defence. *Current Opinion in Immunology* 16(1): 10-15.
- Murphy PM (1993) Molecular mimicry and the generation of host defense protein diversity. *Cell* 72(6): 823-826.

- Obbard DJ, Jiggins FM, Halligan DL, Little TJ (2006) Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Current Biology* 16(6): 580-585.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D *et al.* (2007) Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics* 39(12): 1461-1468.
- Schlenke TA, Begun DJ (2003) Natural selection drives *Drosophila* immune system evolution. *Genetics* 164(4): 1471-1480.
- Silverman N, Maniatis T (2001) NF-kappaB signaling pathways in mammalian and insect innate immunity. *Genes & Development* 15(18): 2321-2342.
- Steiner H (2004) Peptidoglycan recognition proteins: on and off switches for innate immunity. *Immunological Reviews* 198: 83-96.
- Tiffin P, Moeller DA (2006) Molecular evolution of plant immune system genes. *Trends in Genetics* 22(12): 662-670.
- Uttenweiler-Joseph S, Moniatte M, Lagueux M, Van Dorsselaer A, Hoffmann JA *et al.* (1998) Differential display of peptides induced during the immune response of *Drosophila*: a matrix-assisted laser desorption ionization time-of-flight mass spectrometry study. *Proceedings of the National Academy of Sciences of the United States of America* 95(19): 11342-11347.
- Vizioli J, Bulet P, Hoffmann JA, Kafatos FC, Muller HM *et al.* (2001) Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America* 98(22): 12630-12635.
- Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316(5832): 1738-1743.